PATENT ABSTRACTS OF JAPAN

(11)Publication number : 04-228153
(43)Date of publication of application : 18.08.1992

(51)Int.Cl. G11B 20/18
G11B 20/12

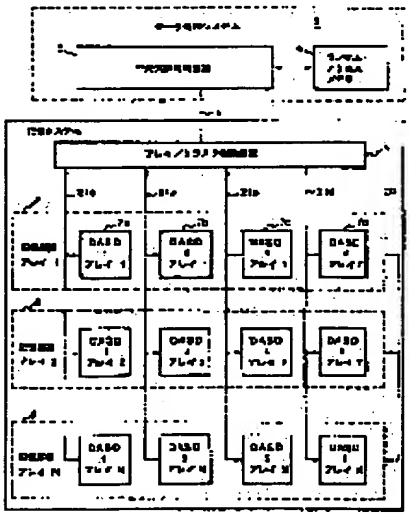
(21)Application number : 03-103708
(22)Date of filing : 14.03.1991

(71)Applicant : INTERNATL BUSINESS MACH CORP <IBM>
(72)Inventor : BEST JOHN S
CHAINER TIMOTHY J
GLASER THOMAS W
GREENBERG RICHARD
MUKHERJEE AVIJIT
NEUBAUER JERRY L
REIDENBACH JOHN R
SCHOPP ROBERT E
SCRANTON ROBERT A

(30)Priority
Priority number : 90 502215 Priority date : 30.03.1990 Priority country : US

(54) DISK STORAGE DEVICE

(57)Abstract:
PURPOSE: To prevent the loss of data even when one driven is out of order by storing data in a first to a (n-1) th drives and parity information in an nth drive.
CONSTITUTION: A disk storage system includes an array/cluster controller 5 and plural storage mechanism arrays 7-9. The constitution of four (n=4) separated disk files 7a-7d into which the respective arrays are divided is preferable since a half/high disk drive is geometrically matched with an existing shape factor of a personal computer. The array is designed as the subsystem of a direct access storage device, is a system so that the data are stored and fetched at high speed and with high reliability and the data are dispersed between the constituent elements of the system.



LEGAL STATUS
[Date of request for examination]
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than

the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号

特開平4－228153

(43)公開日　平成４年(1992) ８月18日

(51)Int.Cl.[5]		識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ１１Ｂ	20/18	１０２	9074－5D		
	20/12		9074－5D		

審査請求　有　　請求項の数11(全 21 頁)

(21)出願番号　特願平3－103708

(22)出願日　平成３年(1991) ３月14日

(31)優先権主張番号　５０２２１５

(32)優先日　1990年３月30日

(33)優先権主張国　米国（US）

(71)出願人　390009531
インターナショナル・ビジネス・マシーンズ・コーポレイション
INTERNATIONAL BUSINESS MASCHINES CORPORATION
アメリカ合衆国10504、ニューヨーク州　アーモンク　（番地なし）

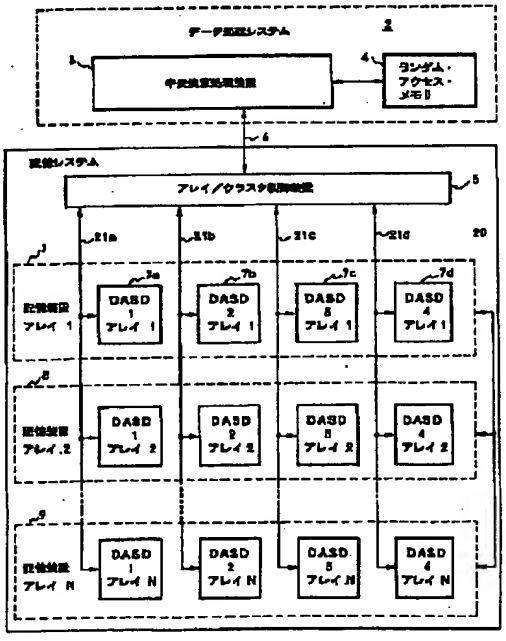
(72)発明者　ジョン・スチュワート・ベスト
アメリカ合衆国カリフオルニア州サンノゼ、オアククレスト・ドライブ　6486番地

(74)代理人　弁理士　頓宮　孝一　（外４名）

最終頁に続く

(54)【発明の名称】　デイスク記憶システム

(57)【要約】
【目的】単一ドライブ・システムよりも信頼性の高い多重ドライブ・ディスク記憶システムを得る。
【構成】Ｎ個のドライブのうち、第１ないし第Ｎ－１番目のドライブにデータを記憶させ、第Ｎ番目のドライブにパリティ情報を記憶させたハード・ディスク記憶システムが提供されるので１つのドライブに故障があったり、遮断されたりしてもデータが失われない。



1

【特許請求の範囲】

【請求項1】（a）磁気記録に適した磁性体のコーティングを有する基板から成るハード・ディスクと、

（b）前記ディスクを回転させるためのスピンドル駆動手段と、

（c）データを読み出し、記録するために前記磁性体と作用結合するトランスジューサ手段と、

（d）前記トランスジューサ手段を、前記ディスク上の複数のデータ・トラックのうちの選択された1本の上方で位置決めするためのアクチュエータ手段と、

（e）N個のドライブとを含み、

データを前記ドライブのうち前記の第1のドライブないし第N－1番目のドライブに記憶させ、前記の第N番目のドライブにパリティ情報を記憶させる、ハード・ディスク記憶システム。

【請求項2】N＝4である、請求項1に記載のハード・ディスク記憶システム。

【請求項3】さらに、ハード・ディスク記憶システムに記憶されるデータ・ブロックを受け取り、前記データ・ブロックをN－1個の部分と1個のパリティ部分に変換し、前記の各部分をそれぞれ前記ディスクの各1枚上に記憶するための制御手段を含む、請求項1に記載のハード・ディスク記憶システム。

【請求項4】N＝4である、請求項3に記載のハード・ディスク記憶システム。

【請求項5】さらに、

（a）前記各スピンドル駆動手段に応答して前記各ディスクの当該の回転位置間の位相差を指示するスピンドル速度誤差信号を発生する手段と、

（b）前記誤差信号に応答して、前記ディスクの回転速度及び相対位置を電子的にロックし、それによって前記の4台のドライブの間にインタリーブされたデータに、読出し及び記録のために同時にアクセスできるようにするスピンドル駆動制御手段とを含む、請求項2に記載のハード・ディスク記憶システム。

【請求項6】さらに、前記の4台のドライブの各々から対応するデータ区画を読み出し、前記区画を最初に記憶されたデータに変換するための制御手段を含む、請求項2に記載のハード・ディスク記憶システム。

【請求項7】前記制御手段が、前記ドライブのうちの3台に記録されたデータ区画、及び前記ドライブのうちの第4のドライブに記録されたパリティ情報に応答して、前記ドライブから読み出された前記データに関するパリティ検査を実行するパリティ検査手段を含む、請求項6に記載のハード・ディスク記憶システム。

【請求項8】さらに、前記ドライブのうちの任意の3台に含まれる対応するデータ・ビットに応答して、最初に記憶されたデータを再生するエラー訂正手段を含む、請求項6に記載のハード・ディスク記憶システム。

【請求項9】さらに、

2

（a）前記の4台のドライブのうちの任意の1台がハード・ディスク記憶システムから遮断されたことを検出し、前記ドライブのうちのどれが遮断されたかを指示する遮断信号を発生する遮断信号手段と、

（b）前記の4台のドライブのうちの任意の1台がハード・ディスク記憶システムに再接続されたことを検出し、前記ドライブのうちのどれが再接続されたかを指示する再接続信号を発生する接続信号手段と、

（c）接続されたままであった前記ドライブのうちの3台から対応するデータ区画を読み出し、前記の遮断されたドライブ上に最初にあったデータ部分を再生し、前記の再生された部分を前記の再接続されたドライブ上に記録する再生制御手段とを含む、請求項4に記載のハード・ディスク記憶システム。

【請求項10】さらに、前記の遮断信号に応答してハード・ディスク記憶システムを動作不能にする手段を含む、請求項9に記載の前記ハード・ディスク記憶システム。

【請求項11】さらに、前記の再接続信号に応答して前記の動作不能にされたシステムを動作可能にする手段を含む、請求項10に記載のハード・ディスク記憶システム。

【発明の詳細な説明】

【0001】

【産業上の利用分野】ディスク記憶装置は、コンピュータで使用する大量のデータ及びプログラムの情報を記憶する問題に対する好ましい手法である。ディスク記憶装置は、30年以上前に登場して以来、より大きな記憶容量、より短いアクチュエータ・アクセス時間、より短い回転待ち時間、及び信頼性を高めるためのディスク故障からの回復を提供するために継続的に改良されてきた。これは、このような装置の占める空間及び消費電力を減らしながら達成されてきた。

【0002】

【従来の技術】ディスク記憶装置の性能の向上は、データ処理システムの他の面の改良に追いつけなかった。最近では、アクセス時間は1／2または1／3に低下し、データ伝送速度は2倍に向上したが、データ処理装置とディスク記憶システムの性能のギャップはますます拡大してきている。これは、大部分は、機械的構成部分をより高速度で移動させるにはより大きなエネルギー量が必要であるという物理的法則の存在によるものである。

【0003】ディスク記憶システムの全体的性能は、典型的なデータ量、たとえば16キロバイト（KB）のデータを取り出すのに必要な時間の量によって特徴づけることができる。最近5年間で、この時間は1／3に低下したが、同じ期間に、システムの半導体部品の性能は10倍に向上した。この不均衡は、改善する兆しがなく、悪化し続けそうである。

【0004】多数の端末にサービスするために通常使用

される中規模データ処理システムに関して性能を見ると、時間の約６０％は記憶サブシステムによって消費されている。通常、１回のユーザ・トランザクションは、約２０回のディスク動作を必要とし、各ディスク動作で約４ＫＢのデータが転送される。このタイプのシステムは、端末操作員への応答時間が１－２秒である。この応答時間は、通常にみられる実際の値であるが、通常の端末動作員にとってこの値は理想的ではない。

【０００５】この応答時間は、記憶サブシステムの性能によって大きく制限され、半導体装置とディスク記憶システムの間の性能ギャップが拡大し続けるにつれてますますそうなる。将来のアプリケーションがますます複雑になり、より多くのデータの転送が必要となると思われるため、この問題の深刻さがさらに増す。ユーザ・トランザクション当たりのディスク動作回数は、２０回から４０回に上昇すると予想され、平均データ転送速度は４ＫＢから２０ＫＢに増加すると思われる。

【０００６】データ処理システムが日常のビジネス及び個人生活のより不可欠な部分になりつつあるため、この問題はさらに複雑になっている。したがって、ユーザが修理のために熟練したサービス担当者を１日程度待つことは、もはや便利でもなく、可能でもない。大抵の場合、システムは、数分あるいはせいぜい１－２時間のうちにサービスできるように復旧しなければならない。そうしなければ、貴重なビジネスの機会が失われる。これは、故障した装置の障害追求及び修理または交換がユーザ自身によって行なわれるようになることを意味する。

【０００７】ディスク記憶装置の回転待ち時間、サイズ、電力消費量、データ伝送速度、ビット密度、トラック密度、保守容易性、及び信頼性という設計上の問題に対する各種の手法がこれまでに用いられてきた。米国特許第３８７６９７８号は、複数の装置をもつ記憶システムを記述している。複数の装置の１つは、パリティ・ビットを記憶するのに利用される。このパリティ・ビットは、対応するビット位置にある当該の各データ用の装置に基づくものである。上記特許の記述によれば、この記憶システムは、１つのカートリッジを使って、他のすべてのカートリッジ内の対応するビットの比較から生ずるパリティ・ビットを表すという、テープ・カートリッジ・システムに関する。一つのカートリッジが失われた場合にそのデータを再生するため、すべてのデータ・カートリッジからすべてのビットを読み出し、パリティ・カートリッジ内の対応するビットの結果を組み合わせて比較し、適正なパリティを生ずるようにしてデータを再生することができる。上記特許で採用された手法は、アクセス機構が１個しかない場合であるが、多重ディスク・ドライブ・システムに使用するのに好適である。ディスク・ドライブ・システムでは、失われたデータを再生するとともに、記憶システムから読み出されたすべてのデータの妥当性のオンライン検査を実行することが望ましい。上記特許に記述された記憶システムを使用しても、ディスク記憶システムから読み出されたデータのオンライン検査に関して満足な性能は得られない。

【０００８】米国特許第４０３６６５９号は、ディスク・ドライブ・システム用のプログラム式制御装置を示している。複数のディスク・ドライブの間でデータを分散することについて長々と詳しく記述されているが、超高速の回転速度で動作する例外的に小さいディスクを使用することは示唆されていない。

【０００９】米国特許第４５７７２４０号に記述されたビデオ記録システムは、別々のアクチュエータを使ってエラー訂正またはディスク・トラブルからの回復の問題を記録するディスク記憶システムに関する。その代わり、他の方法ではエラーを生ずるはずの欠陥を含むトラックは、単純にスキップされる。実際のデータ処理の場では、こうすると、データ容量が容認できないほど失われるはずである。というのは、単一ビット・エラーが１トラック全体の損失をもたらすからである。これは、データ記憶の問題に適用するとき、容認できないトレードオフである。上記特許はまた、複数のドライブ・スピンドルの使用を示唆しておらず、そのかわりに別々のアクチュエータを使用している。この手法は、故障した装置の簡単な交換には役立たない。というのは、２つのアクチュエータ、したがって２つのデータ・グループが関与するからである。

【００１０】１台のディスク記憶システムに複数のスピンドルを使用することは、米国特許第４５８３１３３号で示唆されている。しかし、このシステムでは、一度にただ１つのドライブが使用されるだけである。データをすべての装置の間に同時に配分し記録することができるとの教示はない。ハード・ディスクではなく可換性媒体の使用が意図されており、回転速度が可換性媒体では通常の低速度ではないとの言及はない。

【００１１】米国特許第４７２４４９５号の教示は、連続したビデオ・フィールドをディスク記憶スタックの異なるゾーン上に記録する、別々の２つのアクチュエータを使用したビデオ記録システムを対象とする。別々のスピンドルが使用され、データが複数のスピンドル上に同時に記録されるとの示唆はない。このシステムは、トラック内の欠陥を、そのトラック全体を単にスキップすることによって処理する。これは、上記の米国特許第４５７７２４０号に示された手法に類似している。上記特許は、複数の小型ディスクが、データがすべてのディスクに並列に分散されている１つのシステム内で組み合わされることを示唆していない。

【００１２】このような装置は魅力的な特徴をもつにもかかわらず、まだ克服されていないいくつかの本質的な制限がある。たとえば、データが円形トラックに沿って直列に配列されるため、所望のデータがデータ・トランスジューサの下を通過するまで待つ必要がある。単純な

5

統計的観点からみると、所望のデータがトランスジューサに到達するのに必要な平均時間は、ディスクが1／2回転するのに要する時間である。セクタを1つのトラックから別のトラックへスキューすることによって、この単純な統計的平均待ち時間をやや改善することが可能である。こうすると、トランスジューサは、トラック・アクセス後にこれ以外の方法の場合より多少早くトラックの読出しを始めることができる。データがトランスジューサの下にくるまでの待ち時間は、回転待ち時間と呼ばれる。

【0013】トラックからトラックへのアクセス中にヘッドアーム・アセンブリを機械的に移動させるのに必要な時間は、より高速のアクチュエータを使用して大幅に短縮されてきたが、回転待ち時間はほとんど改良されていない。これは、ディスク・ドライブの仕様を調べると容易に理解される。いわゆる「ハード」ドライブ、すなわち剛体基板を利用した駆動機構は、必ず3600rpmで回転する。回転待ち時間は、回転速度に反比例するので、回転速度を増加させない限り大幅な改良は不可能である。

【0014】スピンドル駆動モータをスピードアップし、それによって回転待ち時間を短縮することは、簡単なことのように思われるが、より高い回転速度のディスク・ドライブがないことから明らかなように、そう簡単ではない。明白な解決には役立たないが、問題のいくつかは明かである。たとえば、ヘッドの空気力学的制御を行なってヘッドとディスクの間隔を20－63．5mμの範囲に維持する、いわゆる「ウィンチェスタ」技術は、毎秒1．5－2．5mの速度で移動する空気の薄い膜の存在を利用している。3600rpmで回転する、直径が約89mm－203mm（3．5－8インチ）以上のディスクでは、この速度要件は容易に満たされ、空気力学的パッケージ内に配置されたトランスジューサである「スライダ」を、これらのパラメータを満たして動作するように設計することができる。

【0015】回転待ち時間は回転速度に直接的に関係し、回転速度を増加させると回転待ち時間の短縮で直接の利益が得られることは以前から認識されてきた。それにもかかわらず、ほぼすべてのハード・ディスク駆動システムの回転速度は、3000－3600rpmの範囲にとどまっている。

【0016】これは、たとえば回転待ち時間を相当に改善するために回転速度を3倍にして約10，000rpmにすると、既存のスライダ技術が全く使えなくなることによって少なくとも部分的に説明される。そのようにすると、増加したディスク速度から生ずる従来とは異なった空気力学的状況に対処するために、コストのかかるスライダの再設計が必要となる。

【0017】空気力学の問題は、予測可能であり、十分な工学的努力によっておそらく解決できるはずである

6

が、他の問題についてはそうはいかない。適切な事前措置が講じられていない場合、ヘッドとディスクの偶然の接触が、装置が寿命に達する前に、故障を生じ得ることが認識されてきた。ディスクの回転開始及び停止時に、トランスジューサを、データが保存されない「着陸」ゾーン上方で位置決めする機構を設計することは可能である。こうすると、データが記憶されるディスク表面の摩耗及び潜在的な損傷の一部分がなくなる。それでもなお、通常のデータ読出し及び書込み中にトランスジューサとディスクの不測の接触が起こる可能性がある。したがって、トランスジューサを着陸ゾーンに位置決めするための特別の機構のコストが経済的に引き合う場合でも、ディスク表面に何らかの形の保護機構を設けることが望ましい。通常、この保護は、ディスク上の記録媒体の表面に塗布される潤滑剤の形をとる。

【0018】適当な潤滑剤の開発は困難であった。必要な潤滑特性の点から、二三のクラスの潤滑剤を除いてすべての潤滑剤が除外される。さらに、潤滑剤は、通常のビジネス環境または家庭環境で見られる汚染物質と干渉すると、通常のヘッドとディスクの間の間隔の実質的部分に取り付く分子、結晶、またはアモルファス構造物を形成するが、このような物理的構造物を形成する汚染物質と相互作用してはならない。ヘッドがこのような構造物と接触すると、「ヘッド・クラッシュ」を起こして、ディスク表面の損傷及びデータの損失をもたらす可能性がある。潤滑剤は、潤滑作用及び化学的性質の点から満足のいくものでなければならないだけでなく、ディスク表面に付着し、回転によって飛び散らない物理的特性ももたなければならない。望ましい化学的特性及び潤滑特性を有するいくつかの潤滑剤は、遠心力の結果ディスクの周辺へ移行する傾向があることが判明している。この障害は、長期間使用後に始めて出現するので、満足のいく性能をもつかどうか潤滑剤を評価し試験するのは困難である。3，600rpmの回転速度で満足なディスク寿命をもたらす働きをする潤滑剤であっても、より高い回転速度で許容できるかどうかを確認するために少くとも長時間の寿命試験が必要となる。このような試験の結果、おそらく既存の潤滑剤はずっと高速の回転速度での使用には満足でないことが示されるはずである。

【0019】したがって、回転速度が3，600rpmに固定され続けたことは理解できる。というのは、より高い回転速度から生じる予見できるが予測不可能な問題が、開発投資の実質的な抑止力になっているからである。

【0020】より高い回転速度を試みた場合、ドライブ内で発散される熱が問題になることが予想される。より高い回転速度でディスクを回転させる場合、必要な電力量が増加するので、必然的にスピンドル駆動モータがより大型になり、それに伴う電力損が増大する。これは、電力損の増大がきわめて小さな容積内で生じ、熱の放散

7

が一層困難になることを認識すれば、小さな問題ではない。また風損が増大し、やはり装置内で発生する熱を増加させることが予想できる。

【0021】占有空間は常に考慮すべき問題なので、より大型のドライブ・モータはより高い回転速度の使用に対するかなりの抑止力になる。ディスク・ドライブがすでに使用可能な空間及び電力のかなりの部分を占めている、いわゆるパーソナル・コンピュータの場合には特にそうである。パーソナル・コンピュータがますます小型化する傾向のもとでは、既存のディスク・ドライブより大きな空間を占めるディスク・ドライブは受け入れられない。さらに、より高い回転速度の装置のヒート・シンクのための空間が余分に必要になると思われる。

【0022】たとえ、回転速度を増加させて回転待ち時間を改善することができたとしても、このような開発には必然的にアクチュエータ・アクセス時間を改善する努力が必要となり、やはり装置内で消費される電力を増加させる傾向にある。

【0023】従来技術は、回転待ち時間を改善する努力とディスク・ドライブに対する他の要件との間に矛盾が存在するという十分な証拠を示している。回転待ち時間を改善した装置の電力とサイズが増大することは、装置をより小型化し、電力消費量を減らし、動作温度をより低くするという要件と矛盾する。とりわけ、装置の信頼性を高め価格を下げることが本質的な要件である。

【0024】従来技術ではうまく対処されなかった別の問題は、欠陥ディスク・ドライブの修理または交換に関するものである。この作業には、従来、平均的なパーソナル・コンピュータ・ユーザの熟練を超える高度の熟練が必要であった。ハード・ディスク・ドライブの修理は、記憶されるデータの性質の故に、精巧度のより低いフレキシブル・ディスク・ドライブよりもクリティカルである。通常、ハード・ディスクは、システムで使用されるすべてのアプリケーション・プログラムを格納し、大量のデータを含むこともある。ハード・ドライブ上の情報はクリティカルなので、ハード・ディスク上のデータをフレキシブル・ディスクまたはテープ型装置上に複製する、「バックアップ」動作を定期的に実行するのが習慣となっている。ハード・ディスクが損傷を受けた場合、またはそれ以外の原因で動作不能になった場合に、フレキシブル・ディスクまたはテープを使ってデータを回復することができる。

【0025】ディスク・ドライブは、誤ったデータに関するエラー訂正が実行できるように設計されているが、このようなエラー訂正能力は、習慣的にきわめて制限されており、小さな規模のヘッド・クラッシュから生ずる大きなデータ・ブロックの損失は処理できない。

【0026】バックアップ動作は実行が面倒であり、バックアップが必要なことは、ハード・ディスク・ドライブが故障するまでしばしば無視される。故障したときは

8

手遅れで、データは回復できない。ディスク・ドライブの交換は、しばしば複雑な作業であり、ディスク・ドライブの取り外しと交換の機械的側面、ならびにディスクをフォーマットし、まだ使用できる情報を再ロードするためのソフトウェア知識において、かなりの専門知識が必要となる。

【0027】従来技術のディスク記憶システムに伴う別の問題は、ディスク基板に使用される材料に関するものである。アルミニウムが、この技術の出現以降ほぼずっと選択されてきた材料である。アルミニウムは、軽量、優れた加工性、及び強度という利点を提供する。以前の値の磁気媒体の厚さ及びヘッド飛行高さでは、アルミニウム基板は、費用はかかるが通常の機械加工技術を使用して適切な仕上げにすることができる。しかしながら、最先端の加工装置を使用しても、アルミニウム基板の表面仕上げは、ビット密度及びトラック密度を向上させるのに必要な薄い媒体コーティングに対する許容差を超える欠陥を含む。

【0028】エラーのないディスクを提供するには、表面仕上げが完全でなければならない。ディスクを磁気媒体でコートしてから、その表面をきわめて微細な表面に仕上げることができる。こうすると、ヘッド・クラッシュに伴う問題の深刻さは軽減されるが、基板内に微細な突起またはピットが存在すると、たとえ媒体の表面がほぼ完全であっても、突起領域内でのビット抜け、及びピット領域内でのビット拡散を生ずる。アルミニウム基板の表面を仕上げるための最先端の技法は、材料の理論的限界に近づいている。純粋なアルミニウムを使用すると、不純物、及びアルミニウム合金に良くみられる包有物を含まないので、良好な仕上げを行なえる可能性が高くなる。残念ながら、純粋アルミニウムはきわめて軟らかく、加工性には乏しいので、得られる表面の品質が制限される。基板に使用するために他の材料も評価した。ガラスと半導体用シリコンが使用されてきた。どちらの材料もアルミニウムよりもずっと良好な表面仕上げを行なえる可能性がある。セラミック基板も使用できる。しかしながら、これらの代替物は、アルミニウムよりも脆い。この特徴のために、これらの材料の広範囲の使用が妨げられてきた。素人による頻繁な移動に耐えられるように比較的頑丈にしなければならないパーソナル・コンピュータで使用されるタイプのドライブでは特にそうであった。さらに、セラミックス、ガラス、及び半導体シリコンの機械的強度は、現在の回転速度で動作の衝撃及び応力に耐えるには十分であるが、回転速度がずっと高くなったときには疑わしい。

【0029】トランスジューサを所望のトラックの上方で位置決めするアクチュエータの性能が絶えず改善されて、アクセス時間がより速くなってきたが、この性能の向上は主として電力消費量の増加という犠牲を払って達成された。これは、電源の容量の増大、バッテリ駆動シ

9

ステムの動作時間の短縮、及びドライブ内で発生する熱の増加の点で、コンピュータ・システム全体にとって問題を生ずる。

【００３０】高性能ディスク・ドライブを使用する大型システムの性能は、各アクチュエータからアクセスできる膨大な量のデータによって課せられる制約に達することが判明している。このような各ドライブ上のデータへのアクセスは必ずしも逐次的には進まないので、データの検査に伴う遅延が避けられない。この制限は、各アクチュエータの下に置くデータを減らすことによって避けることができるが、各アクチュエータ及びそのディスク・ドライブ・システムの関連部品が高価なので、このような手法は経済的に実用的でない。言いかえると、高価なアクチュエータは大量のデータと組み合わせなければ、それを利用する意味がない。高速アクチュエータで比較的少量のデータにアクセスするという目標は、これまで商業的環境では実現不可能であった。

【００３１】さらに、各アクチュエータの下に置くデータを減らすことの利点は認識されてきたが、この問題に対する従来の手法は、追加のアクチュエータの存在により信頼性が失われるという問題があった。

【００３２】要約すると、既存の装置より回転待ち時間を改善し、より小型で、より小さな電力を使用するディスク・ドライブ・システムを実現することが望まれる。この装置は、既存ディスク・ドライブ技術をできるだけ多く利用して既存の潤滑剤及びスライダ技術を使用できるようにするのが理想である。これらの改良は、アクチュエータ性能の向上を伴い、その結果回転待ち時間の短縮を伴うのが理想である。これらのことをすべて、信頼性を低下させずに達成しなければならない。このシステムは、顧客が、複雑なツール、平均的なユーザの能力を越える特別の機械的熟練またはプログラミング能力を使用せずに、ディスク記憶システムの欠陥部分を交換できるようにすることが好ましい。

【００３３】

【発明が解決しようとする課題】本発明の目的は、単一ドライブ・システムよりも信頼性の高い多重ドライブ式ディスク記憶システムを提供することである。

【００３４】本発明の別の目的は、多重ドライブ式ディスク記憶システムを電力遮断する必要なく、かつデータの損失なく、個々のドライブを接続及び切断できる、多重ドライブ式ディスク記憶システムを提供することである。

【００３５】本発明の別の目的は、４つのドライブをもち、そのうちの１つのドライブが故障してもデータの損失を起こさない、または実質的に性能を損なわない、ディスク記憶システムを提供することである。

【００３６】本発明の別の目的は、データの並列検索によってデータがより高速で検索できるように、高速、低コストのアクチュエータを使用して比較的より少量のデータにアクセスすることのできる、ディスク記憶システムを提供することである。

10

【００３７】本発明の別の目的は、アクチュエータ速度の損失またはシステム信頼性の低下を起こさずに、各アクチュエータによってアクセスできるデータの量を減らす、ディスク記憶システムを提供することである。

【００３８】

【課題を解決するための手段】Ｎ個のドライブのうち、第１ないし第Ｎ－１番目のドライブにデータを記憶させ、第Ｎ番目のドライブにパリティ情報を記憶させたハード・ディスク記憶システムが提供されるので、１つのドライブに故障があったり遮断されたりしてもデータが失われない。また、例えば、標準的な１１３ｍｍ（５．２５インチ）のディスク・ドライブに通常必要な空間内にＮ台、例えば４台のユニットを装着することができる。４台のユニットが使用できるので、各ドライブ上に、情報の各バイトの一部分、ならびに４台のドライブのうち任意の３台からデータが再生できるのに十分なエラー訂正情報を割り当て、記憶することができる。こうして、各アクチュエータの下のデータ量がより少ないという利点が、通常なら複数ドライブの使用から生ずるはずのトラブルの頻発なしに、保持される。

【００３９】

【実施例】本発明は、従来技術と著しく異なる機能構成を含む。図１において、ディスク記憶システム１の機能構成は、アレイ／クラスタ制御装置５、及び複数の記憶機構アレイ７、８、９を含む。前記の各アレイは、４つの分離したディスク・ファイル、たとえば７ａ、７ｂ、７ｃ、７ｄを含むことが好ましい。各アレイ内の別々のディスク・ファイルの実際の数は、４より大きくてもよいが、前述したように半高ディスク・ドライブに対する既存のパーソナル・コンピュータ形状因子との幾何的整合性の理由から４が好ましい数である。本発明の構成では、アレイは、直接アクセス記憶装置（ＤＡＳＤ）サブシステム設計である。この設計は、追加の機械的部品をもつより複雑な設計から通常予想されるものとは違って、データの記憶と取出しがより高速かつ高い信頼性で行なわれるような方式で、データをシステムの構成要素の間に分散する。

【００４０】データは、データ・バス６によって中央演算処理装置３に接続されているアレイ／クラスタ制御装置５の監督下で、ディスク記憶サブシステム１とデータ処理システム２の間で受け渡しされる。制御装置５は、データ処理システム１０から受け取って、データ処理システム２のランダム・アクセス・メモリ４に記憶したデータの形式を、アレイにデータを記憶する際に使用する形式に変換する働きをする。制御装置５は、データのバーセルを３つの単位＋パリティ単位にセグメント化し、制御線２０上の信号によって選択されたアレイのドライブａ、ｂ、ｃ、ｄ上に記憶するために、得られた４つの

11

単位を、データ線２１ａ、２１ｂ、２１ｃ、２１ｄを介して転送する働きをする。

【００４１】図２は、データ形式が、データ処理システム２内で使用される形式から、本発明のディスク記憶システムの形式に変換される様式を示す。データ処理システム２のランダム・アクセス・メモリ４に記憶されたデータは、通常、各５２２バイトのパーセル圧縮データ・ブロック１００、１０１ないし１０ｎに配列される。アレイ／クラスタ制御装置５は、図２に示したように動作して、各データ・ブロックを各１７４バイトの３つの区画１００－１、１００－２、１００－３に分割する。区画１００－１、１００－２、１００－３の各々の第１バイトは、バッファ１０１－１、１０１－２、１０１－３のうち対応する１つに置かれる。これらのバッファから、対応するデータ単位は、逐次、線１０３－１、１０３－２、１０３－３を介して対応するバッファ及びエラー訂正コード（ＥＣＣ）生成機構１０５－１、１０５－２、１０５－３に同時に供給される。データ・バッファ１０１－１、１０１－２、１０１－３からの３つの単位は、線１０４－１、１０４－２、１０４－３を介してパリティ生成機構ブロック１１０に供給され、バッファ及びＥＣＣ生成機構１０５－４につながる線１０３－４上に適当なパリティ単位を発生する。

【００４２】線１０６－１、１０６－２、１０６－３、１０６－４上で得られる４つの単位は、データ・ブロックから最初に読み出された３つの対応するデータ・ブロック部分＋１つのパリティ・ビットを表す。これら４つの単位は、４つのＤＡＳＤ装置１００ａ、１００ｂ、１００ｃ、１００ｄの同じ論理アドレスに記録される。各単位は、異なるディスク・ドライブに記録されるが、並列（同時）に記録される。データ・ブロック１００の転送が完了すると、そのデータ・ブロックの第１の単位１００－１はＤＡＳＤ１００－ａに記録され、第２の単位１００－２はＤＡＳＤ１００－ｂに記録され、第３の単位１００－３はＤＡＳＤ１００－ｃに記録され、データ・ブロック１００のパリティ単位はＤＡＳＤ１００－ｄに記録される。

【００４３】記録されるデータは、通常、５２２バイト・ブロックで制御装置５に転送される。５２２バイトのデータ・ブロックの３つの１７４バイト・データ区画への変換は、そのデータ・ブロックの各区画を論理ＦＩＦＯバッファに読み込むことによって実行することが好ましい。このＦＩＦＯバッファは、バッファ１０１－１、１０１－２、１０１－３でよい。

【００４４】欠陥のあるまたは動作しないＤＡＳＤ上で失われたデータを再生する技法が、図３に示されている。この図では、データ部分１００－３を含むＤＡＳＤ１００ｃが、何らかの形で故障して、それに含まれるデータが利用不可能になったと仮定する。これは、図３で、バッファ及びＥＣＣ生成機構１０５－３からデータ・バッファ１０１－３につながる線１０３－３上のＸによって図示されている。ＤＡＳＤの故障を検出するための手段については、後で論じる。ＤＡＳＤ読出し障害が検出されると、システムは図３の配置を取るように再構成される。図３の配置は、図２のそれと本質的に同じであるが、データの流れの方向が逆である。各単位がＤＡＳＤ１００ａ、１００ｂ、及び１００ｄから利用可能になると、バッファ及びＥＣＣ生成機構１０５－１、１０５－２、１０５－４によって、元のデータ形式に変換される。ＤＡＳＤ１００ｃからの単位は入手できない。線１０３－１、１０３－２、１０３－４上の信号が、パリティ・生成機構１１０に供給され、そこで失われた単位が再構築されて、線１０３－３ａを介してデータ・バッファ１０１－３に供給される。線１０３－３ａは、ＤＡＳＤ１００ｃが故障したことをシステムが認識した後で、そのシステムによって行なわれた接続を表す。破線１０３－２ａ及び１０３－１ａは、それぞれＤＡＳＤ１００ｂまたは１００ｃが故障した場合に行なわれる接続を表す。

12

【００４５】好ましい実施例では、データ区画、パリティ単位、及びエラー訂正コードを発生するタスク専用のハードウェアとソフトウェアの組み合わせを利用するが、これらの同じ機能を、適当なプログラムの制御下で動作するマイクロプロセッサ、または完全に専用のハードウェアによって実行することもできる。ほとんどの場合、最適の構成は、専用ハードウェアならびにソフトウェアを含むものである。具体的な実施例は、データ転送速度とマイクロプロセッサの速度の関係、代替各手法の相対的コスト、及びその他の設計上のトレードオフなどの因子に応じて変わる。

【００４６】本発明の好ましい実施例では、５２２バイト・ブロックを３つの１７４単位区画に変換することを企図していることを理解されたい。これらの区画は、次に４つの別々のディスク・ドライブにパリティ単位とともに並列に記録される。これが最適の配置であると考えられるが、他の構成を使用してもよい。たとえば５２２バイト・ブロックを３つの区画に変換する代わりに、各バイトを３つの部分に変換し、または５２２バイトより大きいまたは小さいブロックを使用することも可能である。これらの、またはその他の修正を加える場合、バッファにも修正を加えるのが適切である。最適サイズから外れると、バッファとディスク・ドライブの間のデータの転送が最適より悪くなる結果、性能が低下することがある。他の構成としては、４つより多い、または少ない数のディスク・ドライブを使用することがある。ただし、４という最適のドライブ数から外れると、半高フレキシブル・ディスク・ドライブに対する工業標準との物理的整合性の利点が失われて、最適な４ドライブ構成よりも信頼性が低下することがある。

【００４７】図１に戻ると、アレイ／クラスタ制御装置

13

5は、システム内の各ＤＡＳＤに関する状況情報を維持するという追加の機能をもつ。この機能は、プログラム制御下で動作するマイクロプロセッサによって通常の方法で実行することが好ましい。状況情報は、「ＤＡＳＤリスト」、「性能低下ＤＡＳＤリスト」、及び「ＤＡＳＤ再構築リスト」という３つのデータ・セット内に維持される。ＤＡＳＤリストは、システム内のすべてのＤＡＳＤに関するすべてのパイタル（重要）プロダクト・データを含み、クラスタ内のＤＡＳＤの位置、すなわちＭ行×Ｎ列に編成される。パイタル・プロダクト・データとは、元々ＤＡＳＤ内に含まれ、アレイ／クラスタ制御装置５によってそのＤＡＳＤから取り出される情報である。このような情報は、通常、そのＤＡＳＤの製造連番号、技術変更の数または技術変更レベルあるいはその両方を含む。この情報は、電力投入時及びその他の非常時に、システム内のすべてのＤＡＳＤをポーリングすることにより制御装置によって読み出される。この手順によって、故障して交換されたドライブに関するパイタル・プロダクト・データが、新しいＤＡＳＤを反映するように確実に改訂される。新しいドライブの検出は、単に、各ドライブから読み出されたパイタル・プロダクト・データを、システム内に以前に存在した（ＤＡＳＤの最後のポーリングによって発生された）パイタル・プロダクト・データと比較することによって実行できる。あるＤＡＳＤがアレイ／クラスタ制御装置５から出されたポーリングに応答しない場合、そのＤＡＳＤに、破壊または動作不能のフラグが付けられる。変更されたことが検出されたドライブは、そのドライブに割り振られたデータが再構築されるようにフラグ付けされる。故障したＤＡＳＤ上に以前に含まれていたデータを再構築するための手順については、先に言及したが、後でより詳細に説明する。

【００４８】性能低下ＤＡＳＤリストは、ＤＡＳＤが故障したが、故障した装置を直ちに交換したくないときに使用する。すでに述べたように、ＤＡＳＤが故障しても、システムが完全に動作不能になるのではない。すなわち、故障したＤＡＳＤに含まれていたデータは、図３に関して説明したように再生することができる。故障したＤＡＳＤに含まれていたデータの再生は、欠陥装置の代りに使用される新しいＤＡＳＤ上にそのデータを再書き込みするために行なうこともでき、また再生されたデータを、あたかも故障したＤＡＳＤから直接来たかのように、データ処理システム２が使用することもできる。エラー回復ルーチン後にあるＤＡＳＤがパイタル・プロダクト・データ読出しコマンドまたは書込みコマンドを完了しない場合、そのアレイ内のその特定のＤＡＳＤに性能低下のフラグが付けられる。データ処理システムは、性能が低下した装置に読出しコマンド及び書込みコマンドを出したとき、性能低下ＤＡＳＤリストを利用して、通常エラー回復ルーチンを阻止することができる。

14

このような状況では、図３のデータ回復技法によって、データ処理システムは、転送されたすべてのデータについてエラー回復機能を実行するという負担を課されずに、引き続き機能することができる。もちろん、この能力に対して対価を支払わなければならない。すなわち、同じアレイ内の別のＤＡＳＤが故障すると、そのシステムの動作が継続できないという状況が生ずる。故障したＤＡＳＤをもつシステムの動作は限られた時間しか続けず、故障したＤＡＳＤをできるだけ早く交換することが望ましい。

【００４９】アレイ／クラスタ制御装置５は、故障したＤＡＳＤが動作可能な装置と交換されたことをパイタル・プロダクト・データから認識すると、故障したＤＡＳＤを性能低下ＤＡＳＤリストから削除し、新しいＤＡＳＤをＤＡＳＤ再構築リスト上に置く。このＤＡＳＤ再構築リストは、そのアレイ内の以前に故障した位置に割り当てられていたデータが再生されることを必要とする個々のＤＡＳＤの記録である。ＤＡＳＤ再構築リストは、故障したＤＡＳＤの識別に加えて、データが復元されたＤＡＳＤ内のセクタの数をも含んでいる。セクタ情報を含むリストの使用によって、システムは、データをバックグラウンド・モードで再生することができ、それによって、中央演算処理装置３による通常コマンドの実行が可能になり、データ処理システム２の動作に対する影響を最小にすることができる。

【００５０】アレイ／クラスタ制御プログラムの、図４のデータ流れ図で表される部分では、サブルーチンは、ブロック４０から開始し、システムが電力投入されたことを表す信号、またはシステムがリセットされたことを表す信号を検出する。これらの信号は、通常、データ処理システム内で発生され、さまざまな目的に使用される。

【００５１】ブロック４１で、クラスタ内のすべてのＤＡＳＤをリセットし、そのクラスタ内の各アレイ内の個々のＤＡＳＤをポーリングする。このポーリングによって、そのクラスタ内の各動作可能ＤＡＳＤからパイタル・プロダクト情報が読み出される。各ＤＡＳＤは、その独自の連番号、及び技術変更または技術変更レベルあるいはその両方を示す情報を供給する。さらに、そのＤＡＳＤの個々のセクタに関する情報も提供される。

【００５２】クラスタの各アレイ内のすべての位置のポーリングが完了すると、ブロック４２でテストを実行して、そのクラスタ内のいずれかのＤＡＳＤ位置がそのポーリングに応答しなかったかどうか判定する。あるＤＡＳＤ位置がポーリングに応答しないと、エラー状態と解釈され、プログラムはブロック４２ａで再試行／エラー回復手順に分岐する。この手順は、ブロック４１で表されるポーリング中に適切に応答しなかったＤＡＳＤからパイタル・プロダクト・データを読み出そうと試みる。ブロック４２ａの再試行／エラー回復段階で、前に故障

15

していたＤＡＳＤからバイタル・プロダクト・データを読み出すのに成功した場合、ブロック４３ｂでＮＯ分岐を取って主命令シーケンスに戻る。一方、適切に応答しないＤＡＳＤがまだある場合は、ブロック４３ｂでＹＥＳ経路を取ってプログラムはブロック４２ｃに分岐することができる。ブロック４２ｃで、応答しなかったＤＡＳＤを、それがまだ削除されていない場合、ＤＡＳＤ再構築リストから削除する。さらに、故障したＤＡＳＤを、性能低下ＤＡＳＤリストに追加し、その後、主プログラムに戻る。

【００５３】利用可能なすべての情報が動作可能ＤＡＳＤから読み出され、故障したＤＡＳＤが性能低下ＤＡＳＤリストに追加されると、ブロック４４で、プログラムは現バイタル・プロダクト・データを最後の読出しから得られたデータと比較する。ＤＡＳＤからのバイタル・プロダクト・データは、アレイ内のＤＡＳＤ位置に従って配列される。好ましい実施例では、各アレイ内に４つのＤＡＳＤ、記憶サブシステム内にＮ個のアレイがある。最後の読出し以降に情報が変更された場合、プログラムはＹＥＳ経路を取ってブロック４５ａに分岐し、バイタル・プロダクト・データの変更を反映するようにＤＡＳＤリストを更新する。次にプログラムは、ブロック４５ｂで、故障したＤＡＳＤをＤＡＳＤ再構築リストに追加する。さらに、ブロック４５ｃで、再構築リスト上にあるすべてのＤＡＳＤを性能低下リストから削除することによって、記憶されたＤＡＳＤ情報を更新する。

【００５４】この時点で、ブロック４６で示されるようにサブルーチンは完了し、プログラムはブロック４７で、サブルーチンが完了し、ＣＰＵが記憶サブシステムにデータを転送し記憶サブシステムからデータを要求することができることを示す、「クラスタ・レディー」信号を発生する。次にサブルーチンは、ブロック４８で、次の電力投入またはシステム・リセット信号が検出されるまで、遊休状態に入る。

【００５５】第５図には、システムの電源を遮断しないで欠陥ＤＡＳＤの交換を行なうための技法が示されている。欠陥ＤＡＳＤを交換するためにシステムの電源を遮断する方が簡単であるが、それが望ましくない場合が多数ある。たとえば、システムは、欠陥ＤＡＳＤがあっても、限られた動作をすることができることがあり、電源を遮断すると、この限られた動作が完全に停止することになる。これは、システムがリアル・タイム動作に関係している適用業務、たとえば制御端末やデータ収集システムではきわめて望ましくない。このような環境では、たとえ正常のＤＡＳＤ性能が得られなくても限られた時間動作を継続することがしばしば可能であり、システムを停止するときわめて深刻な結果が生じる。したがって、たとえ限られた時間低い性能で動作することになっても、システムを停止せずに、システムの故障した構成要素の修理または交換が可能な技法が強く求められている。

16

【００５６】あるＤＡＳＤが故障したとき、通常の処置は、できるだけ早く欠陥装置を交換することである。交換は、機械的にもプログラミングの点からもきわめて簡単に実行できるので、コンピュータ保守の熟練者の助けを借りずに、操作員の手でその操作を実行することができる。したがって、訪問サービスを要請し、保守担当員がコンピュータ設置場所に出かけてくるのを待つ遅延がない。こうすると、システムを完全能力に復旧する時間が節約されるだけでなく、訪問サービスの経費も節約される。操作員がシステムの電源を遮断し、欠陥ドライブの物理的交換を実行することも可能かもしれないが、新しいドライブをオンラインにすることも必要である。それには、通常のパーソナル・コンピュータ・ユーザの熟練レベルでは容易に生えられないプログラミング能力が必要であった。

【００５７】したがって、問題は３つある。欠陥ＤＡＳＤの交換が物理的に簡単でなければならず、ＤＡＳＤを交換した後でシステムを動作状態に復旧するのにプログラミングの熟練を必要としてはならず、システムは性能が低下しても欠陥ドライブで動作を継続できなければならない。

【００５８】システムから電源を遮断しないでＤＡＳＤを交換することは「ホット・プラギング」と呼ばれる。ＤＡＳＤがプラグ接続される各アセンブリは、プラグの一部分としてインタロック・ワイヤを含む。ＤＡＳＤをシステムから外すとき、ワイヤが接続されている回路は切断される。システムが外されたＤＡＳＤからデータを読み出そうと試みる場合、すでに説明したように、パリティ回路が失われたデータを再構築する。システムが外されたＤＡＳＤにデータを書き込もうと試みる場合、システムはそのＤＡＳＤに性能低下のフラグを付ける。

【００５９】操作員がシステムをオープンして、欠陥ＤＡＳＤを外すと、インタロック回路がオープンされる。動作可能ＤＡＳＤをプラグ接続すると、インタロック回路が閉じて、割込み信号が発生して制御装置５に送られる。割込み検出は、データ流れ図のブロック５０で行われる。次にプログラムは、ブロック５１に進み、特定のＤＡＳＤクラスタ／アレイ位置に関するバイタル・プロダクト・データ（ＶＰＤ）を読み出す。ブロック５２で、プログラムは、割込みを発生した位置に関するバイタル・プロダクト・データを調べ、それが性能低下ＤＡＳＤリスト上にあったかどうか判定する。ＤＡＳＤ位置が性能低下リスト上にあった場合、プログラムはＹＥＳ分岐を取ってブロック５２ａに進み、そのＤＡＳＤを性能低下ＤＡＳＤリストから削除し、それをＤＡＳＤ再構築リストに追加し、ブロック５４で主プログラムに戻る。

【００６０】ブロック５２で実行されたテストでＮＯの結果となった場合は、そのＤＡＳＤ位置が性能低下ＤＡ

ＳＤリスト上になかったことを示し、ブロック５３に分岐する。ブロック５３で、プログラムは、新しくプラグ接続されたＤＡＳＤからパイタル・プロダクト・データを読み出し、読み出されたデータをそのＤＡＳＤ位置に対して以前に存在したパイタル・プロダクト・データと比較する。パイタル・プロダクト・データが変更されている場合は、新しいＤＡＳＤがその位置にプラグ接続されたことを示し、ＹＥＳ分岐を取ってブロック５３ａに進み、そのＤＡＳＤをＤＡＳＤ再構築リストに追加し、その情報をＤＡＳＤリストに追加し、ブロック５４で主プログラムに戻る。あるいは、ブロック５３のテストでＮＯの結果となった場合は、パイタル・プロダクト・データが変更されていないことを示し、プログラムは単にブロック５４で主プログラムに戻る。

【００６１】元のデータを新しいＤＡＳＤ上で再構築することは、図６のデータ流れ図で表されるプログラムによって実行される。このプログラムの入口は、ブロック６０で表されるアレイ／クラスタ制御装置５プログラム内の遊休ループからである。ブロック６１で、プログラムはＤＡＳＤ再構築リストに照会して、データを復元しなければならないＤＡＳＤがあるかどうかを調べる。このテストは、周期的に、またはアレイ／クラスタ制御装置５が遊休状態にある期間だけ実行することができる。ブロック６１のテストで再構築を必要とするＤＡＳＤがあることが示された場合、ＹＥＳ経路から分岐する。この場合プログラムはブロック６２に移り、ＤＡＳＤ再構築リストを読み出す。もちろん、ブロック６１のテストでＮＯの結果になった場合は、プログラムは、ブロック６３で、単に遊休状態に戻る。

【００６２】ブロック６４で、プログラムは、コマンド割込みを禁止する。これによっては、セクタのブロックの再構築が読出し／書込み動作によって影響を受けないようになる。これは、最も簡単な再構築の方法であるが、他の手法も可能である。たとえば、アレイの残りの部分への通常アクセスを可能にしながら、再構築されるセクタだけをロードする方法がある。このような割込みは、再構築プログラムから出た後に再び許可される。

【００６３】次にブロック６５で、プログラムは、ＤＡＳＤ再構築リスト上の各ＤＡＳＤに対して所定の数のセクタからなるブロックを再構築する。

【００６４】ラウンド・ロビン方式でデータ、すなわち一度にｘ個のセクタの再構築を実行することによって、動作に必要なブックキーピングは大幅に単純化される。さらに、この構築は、システムの全体的動作に最小の影響しか与えないように実行するのが理想的である。主制御プログラムが制御を再確認できる前に１つのＤＡＳＤ全体が再構築される場合、必要な時間の量がシステムの性能に悪影響を与えることがある。ＤＡＳＤ再構築プログラムへの１回のエントリによって再構築できるセクタの数に限度をもうけることによって、システムがこの作業にさく時間の量も同様に制限される。

【００６５】ブロック６６のテストで、再構築の完了したＤＡＳＤがあるかどうか判定する。これは、再構築されたセクタの数を調べることによって判定される。すべてのセクタが再構築されたＤＡＳＤがある場合、ＹＥＳ分岐を取ってブロック６６ａに進み、ＤＡＳＤ再構築リストからそのＤＡＳＤを削除し、ブロック６７で主プログラムに戻る。あるいは、ブロック６６のテストでＮＯの結果となった場合は、プログラムはブロック６７に直接進む。ブロック６７で、ＤＡＳＤ再構築リストに含まれる再構築されるセクタの数を更新する。次にブロック６８で、プログラムは遊休状態に戻る。この状態からブロック６１のテストが再び実行できる。

【００６６】制御プログラムの、ＤＡＳＤアレイからデータを読み出すための部分が、図７に示されている。ブロック７０でサブルーチンに入って、データ処理システム２から読出しコマンドを発行する。ブロック７１で、プログラムは、選択されたアレイ内のすべてのＤＡＳＤにトラック・シーク・コマンドを出す。ブロック７２で、プログラムは、選択されたアレイ内のすべてのドライブ上の所望のデータ・セクタの行を読み出す。ブロック７３のテストで、エラー訂正コードの検査から、またはコマンド・エラーとして、エラー指示があるかどうか判定する。

【００６７】ブロック７３のテストでＮＯの結果になった場合、データ中にエラーがないことを指示し、プログラムはブロック７４に進み、アレイ内のＤＡＳＤから読み出されたデータを組み合わせて、記憶サブシステム１に記憶するためデータ処理システム２から最初に渡された形式でそのデータを置く。

【００６８】しかし、ブロック７３でＹＥＳの結果となった場合は、プログラムはブロック７３ａに分岐し、そこでプログラムはエラー信号を分析して、単一セクタ・エラーが起こったかどうか判定する。そのエラーが単一セクタに限定された場合、ＹＥＳ経路を取って、プログラムはブロック７３ｂに分岐し、そこでエラーのために失われたデータをセクタ・パリティ情報を使って回復する。次にプログラムは、主プログラムに戻り、ＤＡＳＤから取り出されたデータを組み合わせて、最初に記憶されていたデータ・ブロックを再生する。

【００６９】ブロック７３ａからＮＯ経路を取って分岐する場合は、エラーが２つ以上のセクタに広がっていることを示し、プログラムはブロック７３ｃに分岐する。ブロック７３Ｃで、システムは読出し動作を再試行し、さらにデータ回復動作を実行する。これによってうまくエラーの大きさを単一セクタに縮小することができた場合、プログラムはＹＥＳ経路を取ってブロック７３ｄからブロック７３ｂに移る。エラーの大きさが単一セクタより大きいままであった場合は、ブロック７３ｄからＮＯ経路を経てサブルーチンから出て、ブロック７３ｅで

読出し中にハード・エラーがあったと報告する。

【００７０】プログラムが、ブロック７４への経路を取って元のデータを再生する場合、プログラムは続いてブロック７５に進む。ブロック７５で、読出しコマンドをテストして、読み出さなければならないセクタがまだ残っているかどうか調べる。読出し動作がまだ完了していない場合、ＹＥＳ経路を取って、プログラムはブロック７２で読出しプログラムに再び入る。読み出さなければならないセクタがもうなくなった場合は、ＮＯ経路を取って、プログラムはブロック７６に分岐し、読出し動作が首尾よく完了したことを指示する。

【００７１】図８及び図９のデータ流れ図は、システムによって書込み動作がどのように実行されるかを示す。データ処理システム２の中央演算処理装置３が、書込みコマンドをアレイ／クラスタ制御装置５に出す。この書込みコマンドは、メモリ・アドレス、記憶サブシステムに書き込まれるファイルの名前、及びデータが書き込まれるアレイを指定する。

【００７２】ブロック８０でアレイ／クラスタ制御装置５がデータ処理システムから書込みコマンドを受け取ると、プログラムは、ブロック８１に進んで、書込みコマンドによって指定されたアレイを、性能低下ＤＡＳＤリスト上の性能低下ＤＡＳＤのリストと比較する。書込みコマンドによって指定されたアレイが性能低下ＤＡＳＤリスト上にない場合、ＮＯ分岐を取ってブロック８１ａに進み、指定されたＤＡＳＤアレイにトラック・シーク・コマンドを出す。指定されたアレイがこのコマンドを受け取り、アクチュエータが読出し／書込みヘッドをディスク上の希望するトラックの上方に移動させる。シーク・コマンドを出した後、ブロック８１ｂで、図２に示されたデータ・ブロック１００などのデータ・ブロックを取り、それを適当な数、たとえば選択された例では３個の単位（区画）に分割し、その区画に対する新しいパリティ単位を生成する。ブロック８１ｂで実行される動作は、先に図２に関して示し説明した、データ形式の変換及び関連するバッファ記憶に対応する。この動作は、システムで使用されるエラー訂正コードに従ったデータの変更も含んでいた。

【００７３】次にブロック８１ｃで、４つの並列経路を介して選択されたＤＡＳＤアレイ内の４つのドライブに、変換されたデータを書き込む。これは、図２の信号線１０６－１ないし１０６－４によって示される動作である。データの書込みに続いて、ブロック８１ｄで、記録されたデータを検査して、書込みエラーがあったかどうか判定する。これは、ＤＡＳＤから戻された状況コードを調べることによって実行される。トラック・エラーは、書込み動作の前に検出される。エラーが検出された場合、プログラムは、ＹＥＳ経路を取ってブロック８１ｅに分岐し、検出されたエラーの性質に対応するエラー回復手順を実行する。ブロック８１ｆで、書込みエラー

またはシーク・エラーがあるかどうか第２のテストを実行する。ここでエラーが検出された場合は、プログラムは、ＹＥＳ経路を取ってブロック８１ｇに進む　ブロック８１ｇで、ブロック８１ｆで検出されたエラーの追加分析を実行して、２つ以上のＤＡＳＤがエラーを発生したかどうか判定する。ＹＥＳ分岐を取る場合は、２つ以上のＤＡＳＤがエラーを起こしていることを示し、ＹＥＳ分岐を取ってブロック８１ｈに進み、ハード書込みエラー、すなわち同じ位置に再書込みをすることによって訂正されないエラーがあることを指示し、書込み動作が試みられたＤＡＳＤアレイを、性能低下ＤＡＳＤリストに追加する。

【００７４】ブロック８１ｇのテストで単一のＤＡＳＤだけがエラー状態にあることが判明した場合、ＮＯ経路を取ってブロック８１１に分岐し、故障したＤＡＳＤを性能低下ＤＡＳＤリストに追加する。次にプログラムは、ブロック８１ｊに進み、書込みコマンドと、書き込まれたセクタの数を比較して、書き込むべきセクタがまだあるかどうか調べる。セクタがすべて書き込まれている場合、書込みコマンドは完了し、プログラムはＮＯ経路を取ってブロック８２ａに進む。ブロック８２ａで、書込みコマンドが首尾よく完了し、次のコマンドの実行のために記憶サブシステムが使用できることを指示する。

【００７５】ブロック８１ｄに戻って、エラーがあるかどうかのテストでトラック・シーク・エラーまたは書込みエラーが見つからなかった場合、プログラムはＮＯ経路を取ってブロック８１ｋに進み、書込みコマンド中に存在していたセクタの数と、書き込まれたセクタの数を比較し、書き込むべきセクタがまだあるかどうか判定する。ブロック８１ｋからＹＥＳ分岐を取る場合、プログラムは、ブロック８１ｂに戻って、次のデータ・ブロックを取りそれを分割してデータ単位及びパリティ単位を形成し、得られたデータを指定されたＤＡＳＤアレイに書き込む。すべてのセクタが書き込まれている場合、ブロック８１ｋのテストはＮＯ分岐を取ってブロック８１１に進む。このブロックは、ブロック８２ａと類似のものであり、書込みコマンドが首尾よく完了し、次のコマンドの実行のために記憶サブシステムが使用できることを指示する。

【００７６】次にブロック８１に戻って、性能低下ＤＡＳＤリストを書込みコマンドによって指定されたアレイと比較する。指定されたＤＡＳＤがそのリスト上にあった場合、プログラムはＹＥＳ経路を取って、ブロック８３ａに分岐し、プログラムの、データを性能低下モードで書き込む部分に進む。このモードでは、故障ドライブが性能低下ＤＡＳＤリスト上にあることは単に無視される、すなわち故障ＤＡＳＤは、あたかも正しく機能しているかのように取り扱われる。もちろん、これは、通常のエラー回復技法で故障したＤＡＳＤが十分に補償でき

ると仮定したものである。このことについては、後でさらに説明する。

【００７７】ブロック８３で、先に説明したブロック８１ｂと同様に５２２バイトの元のデータ・ブロックを各１７４バイトの３つの区画に分割する。プログラムは、ブロック８３ｃに進み、先に説明したブロック８１ｃと同様に選択されたＤＡＳＤアレイにデータを書き込む。ブロック８３ｃの書込み動作は、あたかも欠陥ＤＡＳＤが実際に適切に機能しているかのように行なわれ、それが正しく機能していないことを無視する。

【００７８】ブロック８３ｄでテストを実行し、書込みエラーまたはトラック・シーク・エラーがあったかどうか判定する。このテストでＮＯの結果になった場合、プログラムはブロック８３ｋでテストを実行し、まだ書き込まれずに残っているセクタがあるかどうか判定する。ブロック８３ｄに戻って、書込みエラーまたはトラック・シーク・エラーが起こった場合、プログラムはブロック８３ｅに分岐し、先に説明したブロック８１ｅと同様にエラー回復手順を実行する。

【００７９】次にブロック８３ｆで、エラー回復手順の結果を検査して、未訂正のエラーがあるかどうか調べる。このブロックは、ブロック８１ｆと同じ手順に従う。すべてのエラーが訂正されている場合、プログラムは、ＮＯ分岐を取ってブロック８３ｋに進み、書込みコマンドによって指定されたセクタを書き込まれたセクタの数と比較する。書き込むべきセクタがまだある場合、プログラムは、ＹＥＳ分岐を取って、ブロック８３ｂに戻る。すべてのセクタが書き込まれている場合、プログラムはＮＯ分岐を取ってブロック８２ａに進む。ブロック８２ａで、書込みコマンドが完了し、別のコマンドを実行するために記憶サブシステムが使用できることを指示する。

【００８０】ブロック８３ｆのエラー回復手順でブロック８３ｄで検出されたエラーから回復できなかった場合、プログラムはＹＥＳ分岐を取ってブロック８３ｇに進む。ブロック８３ｇで、ハード書込みエラー、すなわち回復不能なエラーがあることを指示し、そのエラーを発生したＤＡＳＤを性能低下ＤＡＳＤリストに追加する。次にプログラムは、ブロック８１で主プログラムに戻る。

【００８１】図１０は、個々のＤＡＳＤ装置のスピンドル・モータを制御するためのシステムを示す。すでに説明したように、ＤＡＳＤは、４つのＤＡＳＤのアレイに配置されている。これらのアレイは、任意の数のアレイからなるクラスタに組み合わされる。好ましい実施例では、各アレイは４台のＤＡＳＤから構成される。区分されたデータは、一度に４単位ビットずつアレイに書き込まれる。すなわち、各区画から１単位ずつとそれに対応するパリティ単位（ブロック）が、同時に４枚のディスクに書き込まれる。読出しも並列に実行されるので、ア

レイの４枚のディスクすべての回転位置を、互いに正確な関係に保たなければならない。そうしなかった場合、システム性能は低下する。すなわち、読出し動作は、最後のＤＡＳＤがデータをパスしてしまうまで待たなければならない。したがって、４台のＤＡＳＤが同じ速度で回転し、さらに互いに所定の位置関係を獲得し、記憶サブシステムの動作中ずっとこの関係を維持することが極めて望ましい。

【００８２】この作業の困難さは、１トラック上に記録されるビットの数、すなわちビット密度を考え、各ディスクの位置を１ビット位置の許容差に維持しなければならないことを認識すると理解できる。このレベルの正確さの要件は、各ＤＡＳＤに対して先入れ先出し型の読出しバッファを設けることによって緩和することができる。バッファのサイズは、通常、単一セクタよりわずかに大きいものを扱えるように設定すると好都合である。様々なＤＡＳＤの間の小さなスキューは、データが各バッファから読み出されたときにデータを同期させることによって補正できる。この手法を用いると、スキューの問題が解決され、ＤＡＳＤの回転位置の許容差が緩和されるが、その代りに遅延が増加する。これは、データを使用するとき、最後のディスクがアドレスされたデータを読出しヘッドの下にもってくるまで、システムが待たなければならないためである。通常の状況では、この遅延は、３６００ｒｐｍで動作しているＤＡＳＤで通常見られる回転待ち時間より小さい程度の大きさである。バッファ誘発性遅延は、ディスクが数ビット位置移動するのに必要な時間である。実際の遅延は、ディスクが通常のＤＡＳＤの約３倍の高速で回転していることによってさらに短縮される。

【００８３】図１０で、各ＤＡＳＤ１００ｎは、速度／位置制御機構９０をもつ。これは、図１のアレイ／クラスタ制御装置５内に位置する。各速度／位置制御機構９０は、制御ループ９５をもつ。このループは、ＤＡＳＤからのインデックス・パルスの位置を、加算回路９１内のすべての制御機構９０に対して確定された基準と比較する働きをする。通常の場合、基準インデックス・パルスは、アレイ／クラスタ制御装置５内で水晶制御発振器９２によって発生され、記憶サブシステム及び関連データ処理システムから見てインデックス・パルスの最適な出現時間を表す。

【００８４】インデックス検出器９３は、ディスク９４上のインデックス・マークが通過するごとにインデックス・パルスを供給する。これは、ＤＡＳＤ１００ｎ内のディスクの速度及び位置の尺度を提供する。インデックス検出器９３からのインデックス・パルス及び水晶基準発振器９２からの基準パルスが、加算回路９１で比較され、その結果生じる誤差信号が、制御ループ９５に送られる。好ましい実施例では、ディジタル技術を利用して制御ループ９５を実施するが、制御ループ９５の実際の

実施態様はアナログ、ディジタル、またはハイブリッド技術のいずれでもよい。それが、ＤＡＳＤスピンドル・ドライブ・モータ９７のコイルを励起する、ＤＡＳＤモータ・ドライブ回路９６への信号を発生する働きをすると言うだけで十分である。制御ループは、ＤＡＳＤ１００ｎの速度を、水晶基準発振器９２によって定義される速度に正確に合わせ、インデックス・パルスが正確な時間に丁度インデックス検出器を通過するようにＤＡＳＤ１００ｎのディスクを位置決めすることができる。これによって、すべてのＤＡＳＤの間の位置同期が実現される。あるＤＡＳＤのディスク位置が他のＤＡＳＤのディスク位置よりわずかに後ろにある場合、制御ループ９５は、遅れたＤＡＳＤの速度を、それが正しい速度で回転していても、適切な位置に達するまでわずかに加速させる。適切な位置に達した時点で、他のＤＡＳＤと適合するように速度を下げ、適切な位置を維持する。

【００８５】アレイ／クラスタ制御装置５に使用される電子回路のより包括的な図が、図１１に示されている。中央演算処理装置との交信は、シリアル・インタフェース６を介して行なわれる。このインタフェースは、アレイ／クラスタ制御装置５内にあるシリアル・インタフェース・ドライバ／レシーバ・チップ１２０につながっている。ディジタル制御チップ１２２は、シリアル・インタフェース制御部１２３を含む。この制御部は、アレイ／クラスタ制御装置５とデータ処理システム２の間のシリアル・インタフェースを制御するためのものである。シリアル・インタフェース制御部は、データ処理システム２とアレイ／クラスタ制御装置５の間でやり取りされるすべてのデータ及び制御信号の適切な同期、緩衝及び復号を実行する。

【００８６】サーボ・デモジュレータ・チップ１２７は、読出し線１２７ａからディスク上に書き込まれたアナログ・サーボ・データを読み出し、位置のディジタル表示をバス１２７ｂ上に供給する。

【００８７】ディジタル制御チップ１２２はまた、ディジタル信号処理部１２５を含む。この信号処理部は、サーボ・デモジュレータ・チップ１２７の出力を表すバス１２７ｂ上の読出しヘッド位置データに応答する。バス１２７ｂ上の信号が処理されると、ボイス・コイル・ドライバ・チップ１３０に制御信号が供給される。ドライバ・チップ１３０は、ＤＡＳＤのヘッド・アーム・アセンブリと連動するアクチュエータのボイス・コイル型モータのコイルを励起する。サーボ・デモジュレータ・チップ１２７、ディジタル制御チップ１２２のディジタル信号処理部１２５、及びボイス・コイル・ドライバ・チップ１３０は、読出し／書込みヘッドをアドレスされたトラックの上方で位置決めするためにヘッド・アーム・アセンブリが移動され、データがトラックに書き込まれトラックから読み出されるのに適切な位置に維持されるように、ボイス・コイル・ドライバ・コイルを励起する。

【００８８】さらに、サーボ・デモジュレータ・チップ１２７は、インデックス検出器の機能を果たし、インデックス信号をディジタル信号処理部１２５に渡す。インデックス信号は、先に図１０に関して説明したように水晶基準発振器で比較され、適切な駆動信号がスピンドル・モータ・ドライバ・チップ９６に送られる。ドライバ・チップ９６から、図１０に関して説明したように適切な駆動信号が発生され、スピンドル駆動モータ・コイル９７に送られる。

【００８９】ディジタル制御チップ１２２はさらに、符号化／復号部１２６を含む。この部分は、ＤＡＳＤディスクから読み出されたデータを復号し、そのデータをＤＡＳＤディスク上に記録させる働きをする。記録されるデータは、アナログ波形コンディショナ・チップ１２４に渡される。コンディショナ・チップ１２４は、ディジタル波形を取り、それを適当なアナログ波形に変換して、マルチプレクサ・チップ１２８を介して読出し／書込みヘッド１２８ａ、１２８ｂ、．．．、１２８ｎのうちの選択された１つのヘッドに送る。ＤＡＳＤディスクからデータを読み出すとき、ヘッド１２８ａ、１２８ｂ、．．．、１２８ｎのうちの１つからアナログ信号が選択され、マルチプレクサ・チップ１２８によって増幅され、アナログ波形及びクロック生成機構チップ１２４に渡され、そこで調整されディジタル化される。位相ロック・クロック信号が、読み出されたデータから取り出され、ディジタル化された読出しデータと共にディジタル制御チップ１２２の符号化／復号部１２６に供給される。次に符号化／復号部１２６は、それをディジタル波形に変換し、ディジタル制御装置チップ１２２のシリアル・インタフェース制御部１２３へ転送する。

【００９０】本発明の電気的態様及びシステム態様について説明したので、次に図１２図１３及び図１４に移る。図１２は、記憶サブシステム内で使用される１つのＤＡＳＤの平面図である。図１３及び図１４は、図１２に示したＤＡＳＤの線ＸＩＩ－ＸＩＩに沿って切りとった断面図である。可能な場合、説明される要素を両方の図に示し、同じ参照番号で識別する。ベース・プレート１５０とカバー１５１は、組み合わさって、出来た格納機構内に位置するＤＡＳＤの各種構成要素のシールを行なう。

【００９１】公称直径６５ｍｍの複数のディスク１６０ａ、１６０ｂ、１６０ｃ、１６０ｄが、スピンドル・アセンブリ１６２上に回転できるように配置され、その一端はベース・プレート１５０で、他端はカバー１５１で支持されている。スピンドル・アセンブリ１６２の一部分として、スピンドル駆動モータ１６４、及びディスク１６０ａ、１６０ｂ、１６０ｃ、１６０ｄを適切な位置に保持するためのスペーサ２１３とクランプ２１２が組み込まれている。スピンドル・アセンブリ１６２につい

25

ては、後で図１７を参照して詳しく説明する。

【００９２】ディスク１６０ａ－１６０ｄの各表面は、データを記録するために適当な磁性記録材料でコートされている。図１６に示したように、複数の磁気トランスジューサ１７０ａ及び１７０ａ’ないし１７０ｄ及び１７０ｄ’が、データを読み出し書き込むため、ディスク１６０ａないし１６０ｄの上下のコートされた表面と共同作用するように位置決めされている。記録ヘッド１７０は、当該のサスペンション・アセンブリ１７１ａ及び１７１ａ’ないし１７１ｄ及び１７１ｄ’の端部に装着されている。ヘッド１７０のサスペンションは、通常のジンバル設計に従い、システムの動作中にヘッドをディスク表面から約６．５マイクロインチ（０．１６５ミクロン）浮かせることができる。

【００９３】再び図１２を参照すると、各ＤＡＳＤアセンブリは、ＤＡＳＤへのすべての必要な電力線及び信号線を担持するコネクタ１６９を含む。

【００９４】図１５を参照すると、各ヘッド／サスペンション・アセンブリは、コーム・アセンブリ１７３の対応するアーム１７３ａ－１７３ｅに固定されている。コーム・アセンブリ１７３は、ベアリング・カートリッジ１７５の周りを回転するように配置されている。ベアリング・シャフト端部１７５ａ及び１７５ｂは、ベアリング１７５を両端支持するように、ベース・プレート１５０及びカバー１５１を貫通してそれに結合し、カバー１５１をベース・プレート１５０と係合した状態に保持する助けをする。駆動コイル１７６から構成されるアクチュエータ・コイルは、コーム・アセンブリ１７３の、アーム１７３ａ－１７３ｅと対向する端部に装着されている。

【００９５】コイル１７６がいずれかの極性の直流によって励起されると、駆動コイル１７６内の電流が、それぞれカバー１５１とベース１５０に固定された永久磁石アセンブリ１７７ａと１７７ｂの磁界と共同作用して、コーム・アセンブリ１７３をベアリング１７５の周りで回転させ、ヘッド１７０をディスク１６０上の所望のトラックの上方で位置決めし、データの読出し／書込み中にそれらのヘッドをトラック上方の正しい位置に維持する。磁石アセンブリ１７７ａと１７７ｂはそれぞれ１対の永久磁石１７７ａ１と１７７ａ２、１７７ｂ１と１７７ｂ２を含み、各対はそれぞれ軟磁性体１７７ａ３と１７７ａ４のベース部材上に装着されている。これらのベース部材は、駆動コイル１７６から見た磁気回路を完成し、磁気アクチュエータの駆動効率を向上する低磁気抵抗経路を提供する。１対のクラッシュ・ストップ１７８ａ及び１７８ｂは、それぞれその中点にエラストマ・クッション１７８ｃをもち、コームが許容範囲を超えて移動するのを防止する働きをする。フレキシブル・ケーブル・アセンブリ１７９は、磁気ヘッド１７０及び駆動コイル１７６への必要な接続を行なう。

26

【００９６】図１７に示したスピンドル・アセンブリ１６２は、シャフト１９０を含み、その両端は、ベース・プレート１５０及びカバー１５１中に設けられた対応する穴と対合し、シャフトを両端で支持するようにそれらに結合されている。シャフトを両端で支持することによって、いわゆる「タワー」装着に付随する傾斜及び振動の問題は完全に回避され、完成したアセンブリは振動及び共鳴がほとんどない。シャフト１９０に取り付けられた１対のベアリング１９１と１９２は、ディスク駆動モータの極１９３を支持し、それを固定されたシャフト１９０の周りで回転させる。

【００９７】シャフト１９０に取り付けられた１組の巻線１９４が駆動電流によって励起され、回転磁界を発生する。この回転磁界は、鋼鉄の極１９３の内面に装着された永久磁石２００と共同作用して、その極を回転させる。鋼鉄のスペーサ２０１はベアリング１９２と係合し、極１９３の下部を支持する。極１９３の外面に取り付けられたアルミニウム・ハブ２１０は、その外径がディスク２２０の内径よりわずかに、熱膨張の余地を与える。ディスク２２０及びスペーサ２１３は、シャフト１９０の周りで心合せされ、クランプ２１２によってハブ２１０上で定位置に保持されている。

【００９８】図１８は、４台のＤＡＳＤ２３０ａ、２３０ｂ、２３０ｃ、２３０ｄが単一のプラグ接続可能なカード２３１上にどのように装着されるかを示す。すなわち、このような各カードは、単一のアレイを構成し、複数のこのようなカードがクラスタを形成する。このカードは、通常のプリント回路素材から作成され、接触舌２３２から個々のＤＡＳＤに延びるランド（図示せず）を含む。個々のＤＡＳＤ２３０はそれぞれ、プラグ２３５ａ－２３５ｄを介してカードに接続される。

【００９９】図１９は、ＤＡＳＤ２３０がどのようにカード２３１に固定されるかを示す。各ＤＡＳＤ位置は、キーホール・スロット２３８ａ、２３８ｂ、及び第３のスロット（図示せず）を含む。これらのスロットは、ＤＡＳＤ２３０上の対応するショック・マウント２４０ａ、２４０ｂ、及び第３のマウント（図示せず）、それらに関連するプラグ２３４ｂ１、２３４ｂ２、及び第３のプラグ（図示せず）と一列に並ぶ。

【０１００】図２０は、ＤＡＳＤプラグ及びショック・マウントの配置をより詳細に示す。ＤＡＳＤはプリント回路カード２３１と係合して保持されるが、ショック・マウントによってカードに関連するショックから機械的に隔離される。各ＤＡＳＤは、３つのショック・マウントを含み、それぞれが真中を通るフェルール・ピン２４２、及びエラストマ・グロメット２４０をベース１５０に保持するネジ部分２４３をもつ。ショック・マウント型グロメットは、通常の形状をもち、プリント回路カードの穴の肩を収容する環状のスロットを周囲に含む。グロメットをカード内の鍵形スロットの大きい部分に差し

込むことによって、ＤＡＳＤをカード２３１上に装着する。次にＤＡＳＤを移動させて、グロメット・スロットを鍵形スロットの狭い部分に押し込む。プラグ２３４を鍵形スロットの大きい部分に差し込むことによって、ＤＡＳＤはスロットの狭い部分に保持される。

【０１０１】いま説明したＤＡＳＤの機械的装着は、比較的に簡単であり、使いやすい配置を提供するだけでなく、ＤＡＳＤとカードの間の多数の電気的接触に関する要件をも満たす。システムのプラグ接続可能な各構成要素間の機械的隔離と電気的接触というこの二重の要件は、従来は一般に、固定されたショック・マウントとプラグ接続可能なケーブルによって満たされてきた。この手法は、ほとんどのパーソナル・コンピュータ・システムでいわゆるハード・ドライブの装着の際にとられている。この二重要件のため、たとえば機械的装着の完了後に必要なケーブル接続ができないなど、エラーの機会が増大する。これは、故障を直すためにコンピュータを部分的に分解しなければならないことを意味する。さらに、ケーブルは配線を誤りやすく、システムのコストを増加させる。

【０１０２】図２０に示すように、ＤＡＳＤ２３０のプラグ接続可能なコネクタ１６９は、ソケット２４１につながるアパーチャと隣接するが、それから離れた位置にくるように構成されている。ＤＡＳＤを移動させて、ショック・マウントをキーホール・スロットの狭い部分に位置決めすると、コネクタ１６９は、コネクタ１６９上のランドがソケット２４１内の対応する接点と接触するように、ソケット２４１のアパーチャ内に入る。従来のソケットは、ＤＡＳＤとカード２３１の間の機械的隔離をもたらさないので、新規な配置が提供される。コネクタ１６９のランドとカード２３１上の対応するランド、たとえばランド２４５及び２５６との間の電気的接続は、バネ２４７及び２４８によって実施される。挿入中のソケットの位置決めは、２本のピン２６０によって行なわれる。これらのピンは保持用座金２６０ａを含み、これらの座金は、カード２３１の過大なサイズの穴中でピンを浮かせる。これは、ＤＡＳＤコネクタ１６９の挿入中の過剰な移動を防止するが、その後係合が外れて、ＤＡＳＤ２３０をカード２３１から機械的に隔離する。

【０１０３】本発明の利点は、交換の前にシステムの電源を遮断し、交換後にシステムの電力を再度投入する必要なしに、故障したＤＡＳＤを取り外し、それを動作可能な装置と交換できることである。したがって、この動作モードは、「ホット・プラギング」と呼ばれる。図２１で、ホット・プラギング機能の機械的態様を図示する。各ＤＡＳＤアレイ３００は、信号及び電源用の複数の接触ランドを含むプラグ接触部分３０２をもつカード３０１を備えている。プラグ接触部分３０２上のランドは、データ処理システムのフレームに取り付けられたソケット３１０の対応する接点と共同作用する。ＤＡＳＤ

がソケット３１０で電源にプラグ接続できるように、各ランド３２０は電源ランド３２１及び３２２よりも物理的に短くなっている。このため、ＤＡＳＤ３００内の回路及びアレイ／クラスタ制御装置５内の回路がプラグ接続操作中に優位な信号を受け取らないことが保証される。信号回路と電源回路が混合方式でまたは同時に接続される場合、それらの回路はスプリアス信号を発生し、アレイ／クラスタ制御装置に、またはさらに悪くするとデータ処理システムにエラーをもたらす。信号接続が行なわれる前にＤＡＳＤ３００の回路に適切に電力を供給することによって、スプリアス信号の発生が回避される。プラグ３０２上のランドの機械的配置は、電源ランドが接続されるまで信号ランドへの接触を物理的に防止する。挿入中、電源接触を行なってから、信号接触を行なうまでの間の時間遅延により、すべての回路に適切に電力が供給され、それによってエラー信号の発生が防止される。

【０１０４】

【発明の効果】

【０１０５】本発明によれば、単一ドライブ・システムよりも信頼性の高い多重ドライブ・ディスク記憶システムが得られる。

【０１０６】本発明によれば、多重ドライブ・ディスク記憶システムを電力遮断する必要なく、かつデータの損失なく、個々のドライブを接続及び切断できる、多重ドライブ・ディスク記憶システムが得られる。

【０１０７】本発明によれば、４つのドライブをもち、そのうちの１つのドライブが故障してもデータの損失を起こさない、または実質的に性能を損なわない、ディスク記憶システムが得られる。

【０１０８】本発明によれば、データの並列検索によってデータがより高速で検索できるように、高速、低コストのアクチュエータを使用して比較的により少量のデータにアクセスすることのできる、ディスク記憶システムが得られる。

【０１０９】本発明によれば、アクチュエータ速度の損失またはシステム信頼性の低下を受けずに、各アクチュエータによってアクセスできるデータの量を減らす、ディスク記憶システムが得られる。

【図面の簡単な説明】

【図１】本発明のディスク記憶システムのアーキテクチャを示す機能ブロック図である。

【図２】データの変換、及びデータ処理システムから記憶サブシステムへのデータの転送を示す機能ブロック図である。

【図３】ディスク・ドライブが動作不能になった条件下での記憶サブシステムからデータ処理システムへのデータの転送を示す機能ブロック図である。

【図４】システム電源投入及びシステム・リセット中にアレイ／クラスタ制御装置が利用するプログラムを示す

データ流れ図である。

【図５】制御プログラムの、システムの電源を遮断せずに欠陥ＤＡＳＤの取外し及び正常ＤＡＳＤの代替使用を行なう部分を示すデータ流れ図である。

【図６】アレイ／クラスタ・プログラムの、欠陥があると判定されたＤＡＳＤの代りにに新しいＤＡＳＤ上にデータを再構築する部分を示すデータ流れ図である。

【図７】プログラムの、ＤＡＳＤから読み出されたデータを取り、パリティ検査し、元のデータを再生する部分を示すデータ流れ図である。

【図８】プログラムの、データ処理システムからのデータをＤＡＳＤサブシステムに書き込む部分を示す左半分のデータ流れ図である。

【図９】プログラムの、データ処理システムからのデータをＤＡＳＤサブシステムに書き込む部分を示す右半分のデータ流れ図である。

【図１０】プログラムの、スピンドル駆動モータを制御するのに使用される部分を示すデータ流れ図である。

【図１１】記憶サブシステム制御装置によって利用されるシステムの電子回路の概略図である。

【図１２】記憶サブシステム制御装置で使用されるディスク・ドライブの平面図である。

【図１３】図１２に示したディスク・ドライブの側面図である。

【図１４】図１２に示したディスク・ドライブの側面図である。

【図１５】アクチュエータ及びヘッド・アーム・アセンブリの展開図である。

【図１６】ヘッド・アーム・アセンブリ及びディスクの両端の部分側面図である。

【図１７】ベアリング支持体の中心で切ったスピンドル駆動モータの側面図である。

【図１８】５．２５インチのディスク・ドライブの標準の形状因子の整数分の１に適合し、個々のディスク・ドライブを切断及び接続することのできる、カード上に装着された４台のディスク・ドライブの配置図である。

【図１９】個々のＤＡＳＤがどのようにカード上に装着されるかを示す図である。

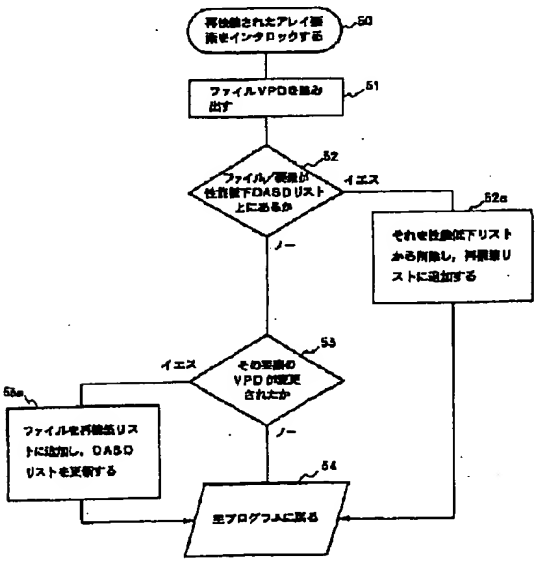
【図２０】ＤＡＳＤ用のショック・マウント及びソケットの詳細な断面図である。

【図２１】個々のクラスタをアレイから外して交換できる、プラグ及びガイドの配置を示す図である。

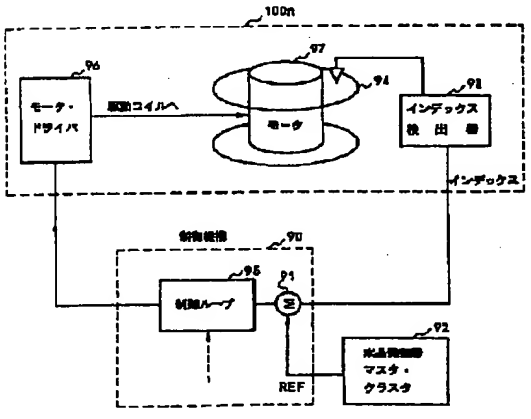
【符号の説明】

1　ディスク記憶システム
2　データ処理システム
3　中央演算処理装置
4　ランダム・アクセス記憶装置（ＲＡＭ）
5　アレイ／クラスタ制御装置
7　記憶装置アレイ
8　記憶装置アレイ
9　記憶装置アレイ
１０　データ処理システム
１０１　データ・バッチ
１０５　バッファ及びＥＣＣ生成機構
１１０　パリティ生成機構

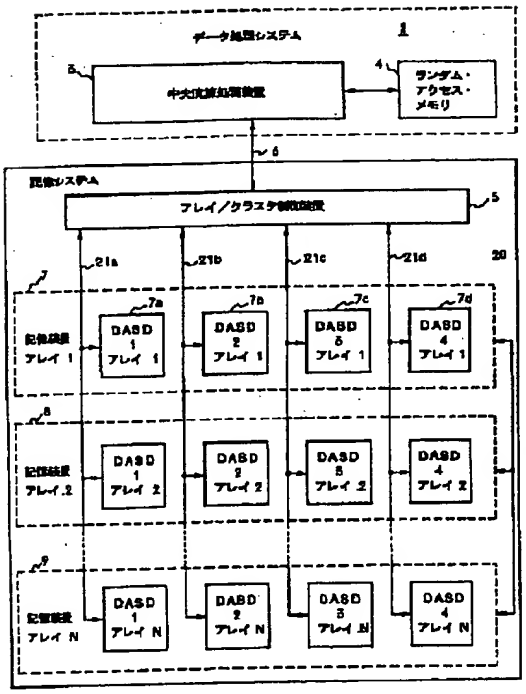
【図５】



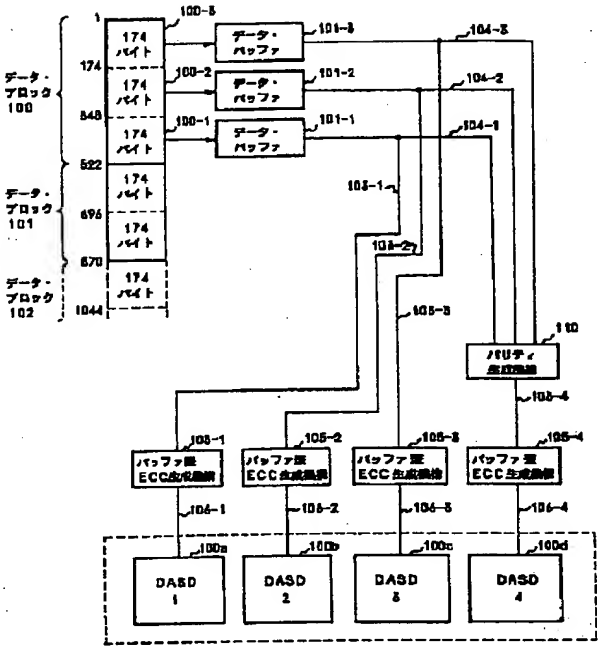
【図１０】



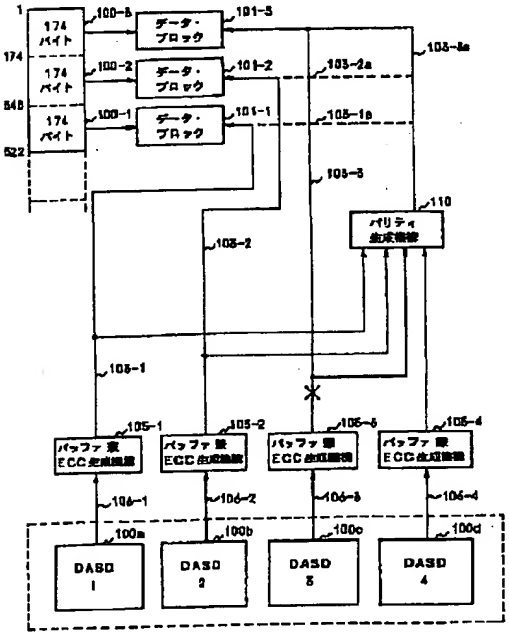
【図1】


データ処理システム
1
中央処理装置
4
ランダム・アクセス・メモリ
6
記憶システム
アレイ/クラスタ制御装置
5
21a
21b
21c
21d
20
7
8
9
記憶装置 アレイ 1
記憶装置 アレイ 2
記憶装置 アレイ N
DASD 1 アレイ 1
DASD 2 アレイ 1
DASD 3 アレイ 1
DASD 4 アレイ 1
DASD 1 アレイ 2
DASD 2 アレイ 2
DASD 3 アレイ 2
DASD 4 アレイ 2
DASD 1 アレイ N
DASD 2 アレイ N
DASD 3 アレイ N
DASD 4 アレイ N

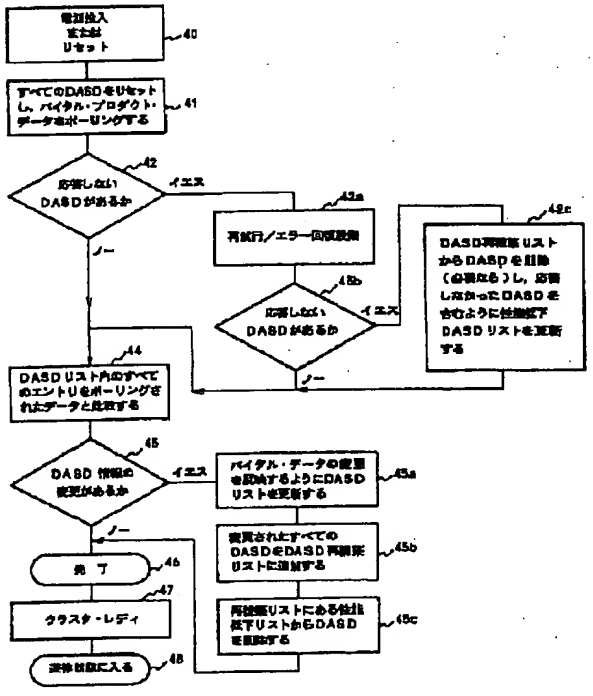
【図2】


データ・ブロック 100
データ・ブロック 101
データ・ブロック 102
174 バイト
174
348
522
696
870
1044
100-3
100-2
100-1
データ・バッファ
101-3
101-2
101-1
104-3
104-2
104-1
103-1
103-2
103-3
パリティ生成回路
110
103-4
バッファ型 ECC生成回路
105-1
105-2
105-3
105-4
106-1
106-2
106-3
106-4
100a
100b
100c
100d
DASD 1
DASD 2
DASD 3
DASD 4

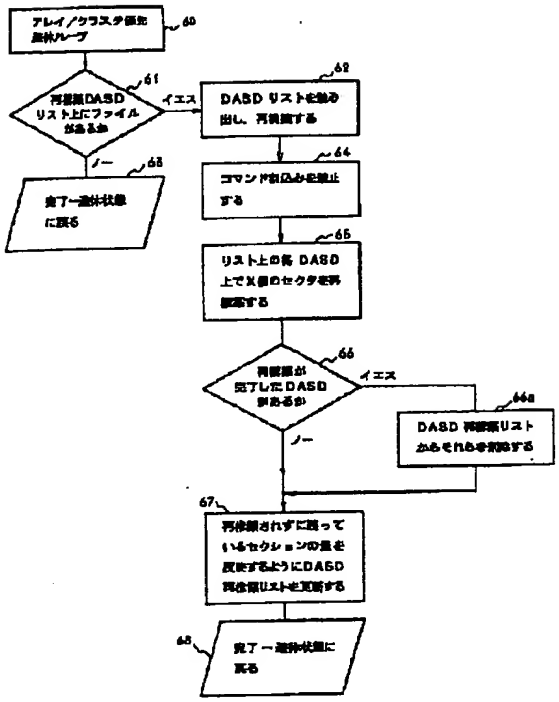
【図3】


1
174 バイト
174
348
522
100-3
100-2
100-1
データ・ブロック
101-3
101-2
101-1
103-3a
103-2a
103-1a
103-3
110
パリティ生成回路
103-2
103-1
105-1
105-2
105-3
105-4
バッファ型 ECC生成回路
106-1
106-2
106-3
106-4
100a
100b
100c
100d
DASD 1
DASD 2
DASD 3
DASD 4

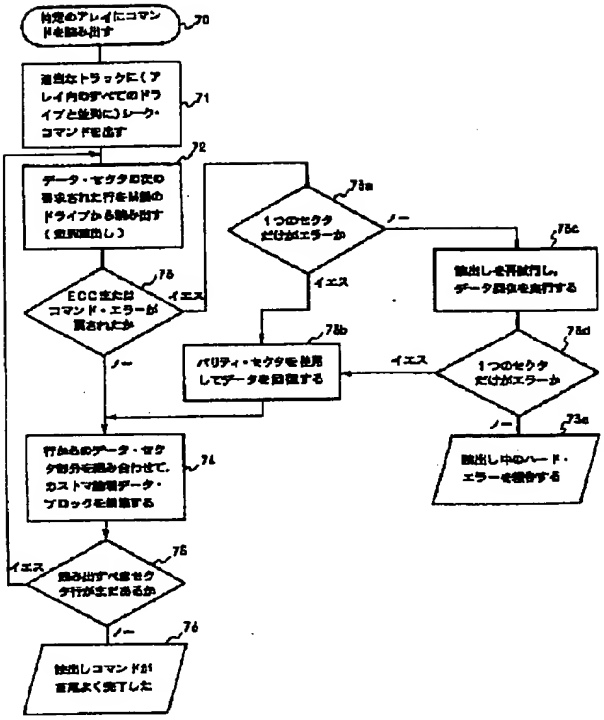
【図4】


40
41
すべてのDASDをリセットし、バイタル・プロダクト・データをポーリングする
42
応答しないDASDがあるか
イエス
ノー
42a
再試行/エラー回復処理
42b
応答しないDASDがあるか
イエス
ノー
42c
DASD再構築リストからDASDを削除（必要なら）し、応答しなかったDASDを含むように性能低下DASDリストを更新する
44
DASDリスト内のすべてのエントリをポーリングされたデータと比較する
45
DASD情報の変更があるか
イエス
ノー
45a
バイタル・データの変更を反映するようにDASDリストを更新する
45b
変更されたすべてのDASDをDASD再構築リストに追加する
45c
再構築リストにある性能低下リストからDASDを削除する
46
完了
47
クラスタ・レディ
48

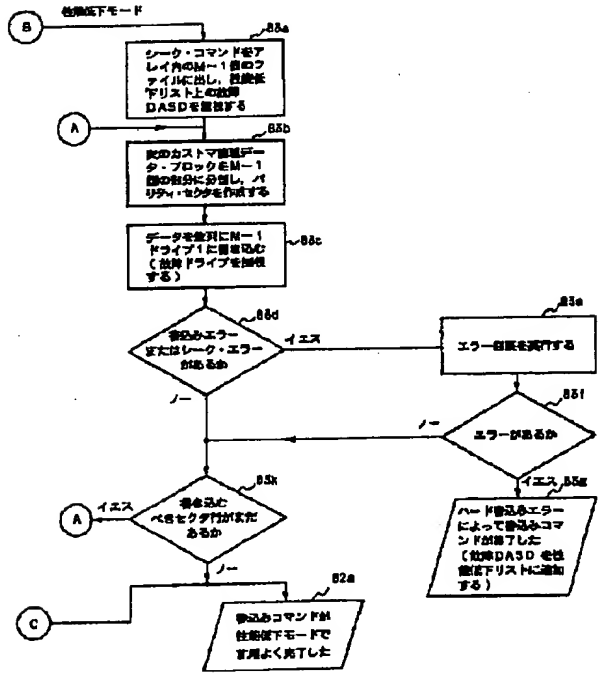
【図６】


アレイ/クラスタ固定
無休ループ
60
再構築DASD
リスト上にファイル
があるか
61
イエス
DASD リストを読み
出し、再構築する
62
ノー
63
完了－通常状態
に戻る
コマンド組込みを禁止
する
64
リスト上の各 DASD
上でX個のセクタを再
構築する
65
再構築が
完了した DASD
があるか
66
イエス
DASD 再構築リスト
からそれらを削除する
66a
ノー
67
再構築されずに残って
いるセクションの量を
反映するようにDASD
再構築リストを更新する
完了－通常状態に
戻る
68

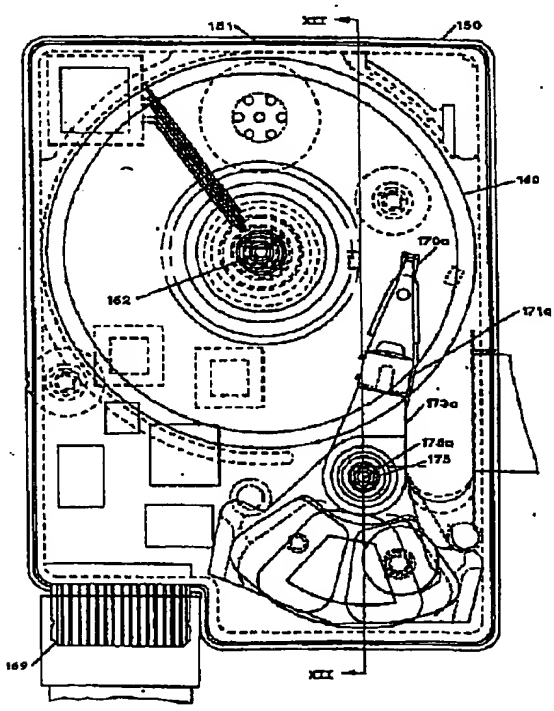
【図７】


特定のアレイにコマン
ドを読み出す
70
適当なトラックに（アレイ内のすべてのドライブと並列に）シーク・コマンドを出す
71
データ・セクタおよび
書き込まれた行を残りの
ドライブから読み出す
（並列読出し）
72
1つのセクタ
だけがエラーか
73a
ノー
イエス
73c
読出しを再試行し、
データ回復を実行する
ECCまたは
コマンド・エラーが
見られたか
75
イエス
ノー
パリティ・セクタを使用
してデータを回復する
73b
イエス
1つのセクタ
だけがエラーか
73d
ノー
読出し中のハード・
エラーを報告する
73e
行からのデータ・セク
タ部分を組み合わせて、
カストマ論理データ・
ブロックを構築する
74
イエス
読み出すべきセクタ行がまだあるか
75
ノー
76
読出しコマンドが
首尾よく完了した

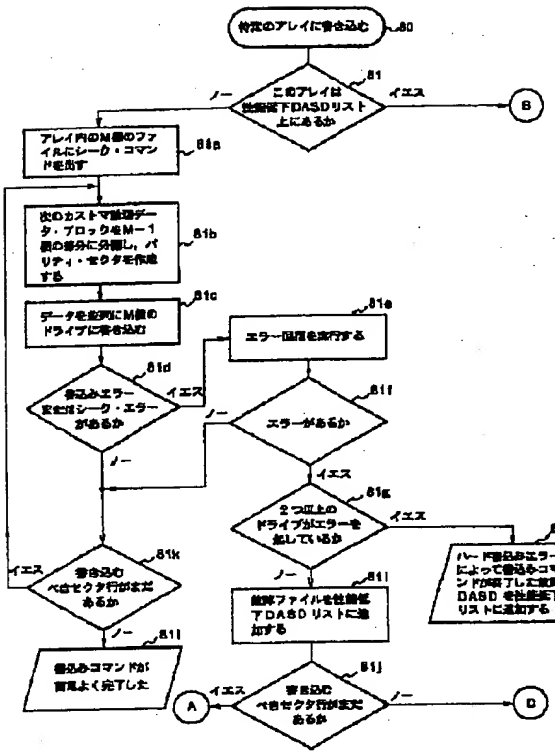
【図９】


B
性能低下モード
シーク・コマンドをア
レイ内のM－1個のフ
ァイルに出し、性能低
下リスト上の故障
DASDを識別する
83a
A
欠けたカストマ論理デー
タ・ブロックをM－1
個の部分に分割し、パ
リティ・セクタを作成する
83b
データを並列にM－1
ドライブに書き込む
（故障ドライブを無視
する）
83c
書込みエラー
またはシーク・エラー
があるか
83d
イエス
エラー回復を実行する
83e
ノー
ノー
エラーがあるか
83f
イエス
83g
ハード書込みエラー
によって書込みコマン
ドが終了した
（故障DASD を性
能低下リストに追加
する）
A
書き込む
べきセクタ行がまだ
あるか
83h
イエス
ノー
C
書込みコマンドが
性能低下モードで
首尾よく完了した
82a

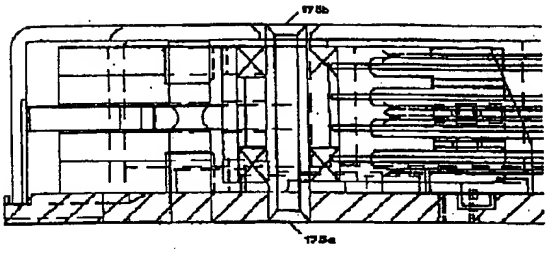
【図１２】


151
XII
150
140
170a
162
171a
173a
175a
175
169
XII

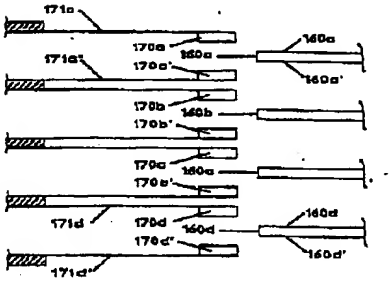
【図８】



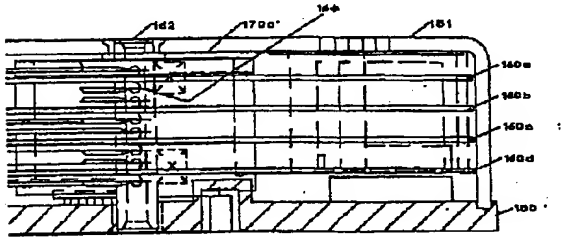
【図１３】



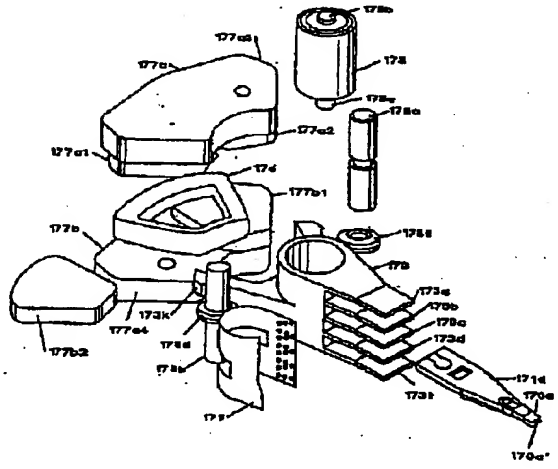
【図１６】



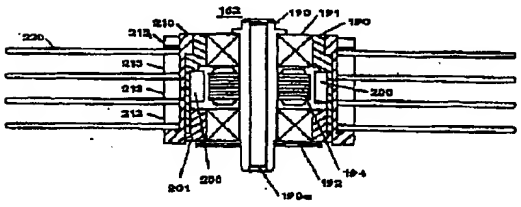
【図１４】



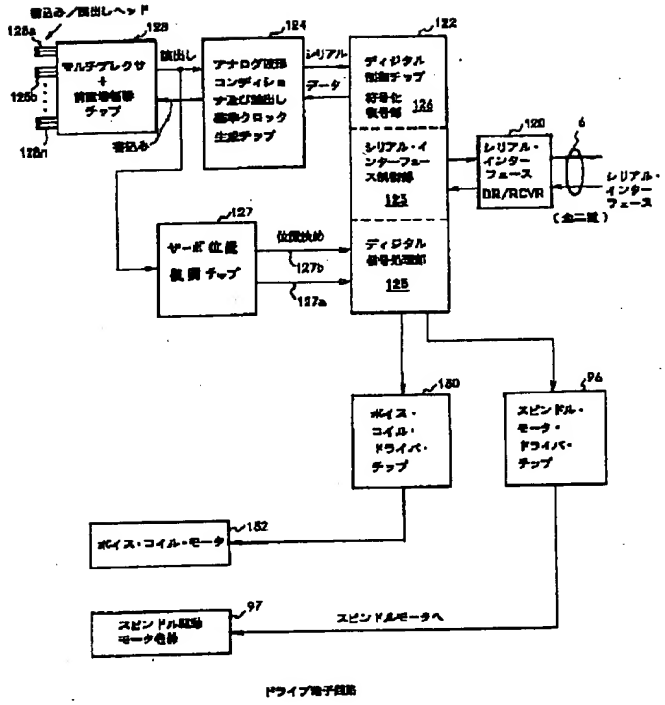
【図１５】



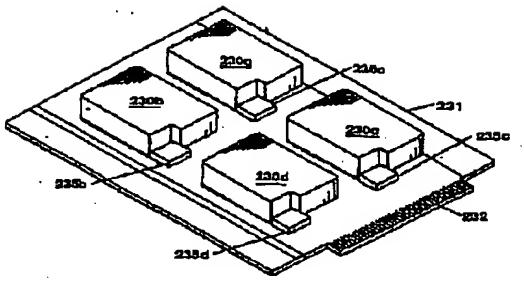
【図１７】



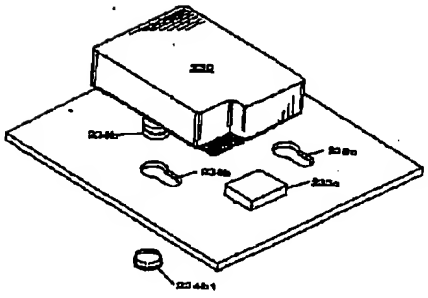
【図１１】



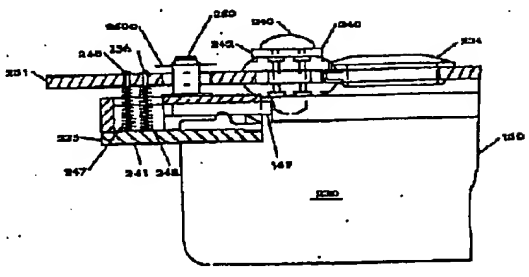
【図１８】



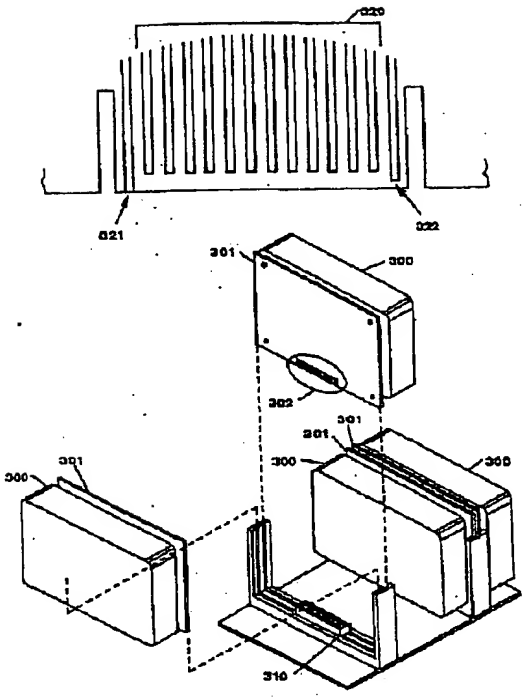
【図１９】



【図２０】



【図21】



フロントページの続き

(72)発明者　テイモシー・ジヨセフ・チヤイナー
アメリカ合衆国ニユーヨーク州マホパツク、アールデイ　ナンバー1　パレツト・ヒル・ロード　（番地なし）
(72)発明者　トーマス・ウイリアム・グラサー
アメリカ合衆国ミネソタ州ロチエスター、ボツクス183、　ルート8番地
(72)発明者　リチヤード・グレンバーグ
アメリカ合衆国ミネソタ州ロチエスター、ノーザン・ヒルズ・ドライブ・ノース・イースト　748番地
(72)発明者　アヴイジツト・ムカジー・ムカジー
アメリカ合衆国カリフオルニア州サンノゼ、アンジユ・クリーク・サークル　7143番地
(72)発明者　ジエリー・リー・ネウバウアー
アメリカ合衆国ミネソタ州ロチエスター、フイフス・アベニユー・サウス・イースト　706番地
(72)発明者　ジヨン・ラルフ・レイデンバツチ
アメリカ合衆国ミネソタ州ロチエスター、フイフス・アベニユー・ノース・ウエスト　3001番地
(72)発明者　ロバート・エルスウオース・スコツプ
アメリカ合衆国ミネソタ州ロチエスター、セブンテイーンス・アベニユー・ノース・ウエスト　2712番地
(72)発明者　ロバート・アレン・スクラントン
アメリカ合衆国カリフオルニア州サンノゼ、ピユウクレスト・コート　20242番地